译 者 的 话

P. 亨利西著《常微分方程离散变量方法》一书出版于1962年，我们当即译出作为计算数学专业高年级专业课的教材，教学效果较好，书中提出的某些问题曾作为学生的毕业论文予以完成. 该书问世至今已近二十年，在此期间，国际上相继出版了不少常微分方程数值解法方面的著作，其中很多是优秀作品.就在这些优秀著作中，在重要理论的叙述和关键公式的推导方面，经常引用 Henrici 的这本书或者建议读者参看这本书. 近年来，国内同行普遍认为将该书译成中文是有价值的，鼓励我们整理初稿，予以出版，以飨读者.

这里，我们特别感谢刘德贵同志，他对译稿进行了详尽的校阅；我们还对王长富同志的校对工作表示感谢.

译 者

1981.2.

序

本书既准备作为大学高年级的教科书，同时还介绍最新的重要理论结果．本书的基础部分着重于叙述基本概念以及有实质性意义的问题，而不是像手册那样列举专门技术的细节，这一点反映了我对讲授数值分析的看法．本书大部分较深的内容(特别是误差传播方面的材料)是新的．

我曾为数学工作者以及爱好数学的工程师和物理学家多次讲过题为“微分方程数值解法”的课程，本书来源于这一课程的讲稿．这一课程所包括的主要内容，大约为现在这本书的第一、二章和第五章的前半部分．对于这些内容，必须具备的预备知识是微分方程的初等教程、高等微积分的某些部分和一些关于计算机计算的知识(在洛杉矶加州大学，这些关于计算机计算的知识，通过选学一门一个学分的初等程序设计课程就可获得)．第三、四章和第七章需要一些矩阵和线性代数的知识．第五及第六章的某些部分需有初等复变函数理论方面的准备知识．

每章最后所附的习题主要是理论性的．显然，一个学生要想完全理解本书中介绍的算法，就必须将它们应用到具体的实际问题上去．然而，能够合理地要求一个学生所求解的问题的大小，在很大程度上取决于可供其使用的计算设备．由于这个原因，对于数值计算的问题除了某些例外都留给教师自行选择．

影响到我对问题的见解的形成的因素是多方面的，这里只能提及几个方面．关于多步法的几章，它们现在的表达形式

很多来源于 Dahlquist [1956,1959] 的两篇重要文章；关于数值稳定性的几节，基于 Rutishauser [1952] 的工作；关于 Runge-Kutta 方法的几节，基于 Gill [1951] 的工作；关于采用的记号、计算的专门知识和过去的参考文献，则主要依据 Collatz [1960] 的工作。

P. 亨利西

1961 年 9 月

加利福尼亚，洛杉矶

目 录

第II部分 初值问题的多步方法

第III部分　边值问题

引　　言

0.1. 定义. 问题的分类

令 $-\infty < a \leqslant \infty$. 我们用 $[a,b]$ 表示满足 $a \leqslant x \leqslant b$ 的所有的实数 x 的集合. 如果 $-\infty \leqslant a \leqslant b \leqslant \infty$, 我们用 (a,b) 表示满足 $a < x < b$ 的所有 x 的集合. 符号 $(a,b]$ 和 $[a,b)$ 按照类似方式来使用. 令 p 是一个整数且 $p \geqslant 1$, 又令 $a < b$ 且 $F(x, y_0, y_1, \cdots, y_p)$ 是定义在 $x \in [a, b]$ 和 $(y_0, y_1, \cdots, y_p) \in D$ 中的实函数, 其中 D 是实 $(p+1)$-维 Euclid 空间的一个区域. 方程

$$F(x, y, y', \cdots, y^{(p)}) = 0 \tag{0-1}$$

称为 p 阶微分方程. 若函数 $y(x)$ 有定义且在 $[a, b]$ 的一个子区间 I 上 p 次可微, 当 $x \in I$ 时, 点 $(y(x), y'(x), \cdots, y^{(p)}(x))$ 在 D 中, 并使得

$$F(x, y(x), y'(x), \cdots, y^{(p)}(x)) = 0, \quad x \in I \tag{0-2}$$

成立, 则称 $y(x)$ 为微分方程的一个解. 从简单的例子便可知道, 一个给定的微分方程, 可以有许多解. 例如, 若 $p = 1$, 而且 $F(x, y_0, y_1) = y_0 - y_1$, 则每一个函数 $y(x) = Ce^x$ (C=常数)为一个解. 为了确定给定的微分方程的一个解, 通常需要说明它的某些附加性质, 例如在指定的点上给出函数值或它的导数值. 结果是: 一个 p 阶方程一般恰好需要 p 个定解条件. 一个重要的特殊情况是以条件

$$\begin{aligned} y(a) &= \eta_0, \\ y'(a) &= \eta_1, \end{aligned} \tag{0-3}$$

$$\cdots\cdots$$

$$y^{(p-1)}(a)=\eta_{p-1}$$

所表示的，其中 $\eta_0, \eta_1, \cdots, \eta_{p-1}$ 为已知常数．确定 $y(x)$ 使它满足 (0-2) 和 (0-3) 的问题称为初值问题．边值问题是这样的问题，函数 $y(x)$ 以及（或）它的某些导数在几个不同的点上取指定的值．

当 $p=1$ 时是最简单的初值问题．通常假设 (0-1) 已就最高阶导数解出，于是一阶初值问题形如

$$y'=f(x,y),\quad y(a)=\eta, \tag{0-4}$$

其中 η 为一常数．为了避免较烦的讨论，我们将总假设函数 $f(x,y)$ 是对 $x\in[a,b]$ 以及对所有有限的 y 都有定义．在第一章、第二章及第六章里将研究这类初值问题，第三章考虑一阶方程组的初值问题，第四章和第六章考虑阶 >1 的方程的初值问题，第七章则为处理二阶方程的边值问题．

0.2. 求解微分方程的数值方法的必要性

例如 (0-4) 这样一个初值问题，它的解在数学上的存在性，对于 $f(x,y)$ 来说，可以在非常一般的条件下得到证明．第一章给出了基本存在定理的概述，而在第三章叙述方程组的情形．然而在微分方程的许多应用中，所要求的不仅是解在数学上的存在性——这一点常常可以从非数学的考虑得到证明——而且是解在自变量的指定范围内取值时它的（近似）数值．关于这方面的情况是很少令人满意的．对某些实际上是重要的但却是十分特殊的微分方程类，它的解可用封闭形式给出，即以初等函数诸如多项式、指数函数、对数函数以及这些函数的不定积分的有限组合给出．另一方面，许多其它的微分方程，并不能按照这种方式来求解，即使外形看来是简单的微分方程，例如

$$y' = x^2 + y^2 \text{ 或 } y'' = 6y^2 + x.$$

已经证明它们的解不能以初等函数来表示．的确，虽然有相当一类具有显式解的微分方程(见 Kamke [1943])，但可以肯定地说，大多数微分方程不能求得其显式解．必须重视的是，即使在显式解存在的情形下，寻找它的数值解的问题也不一定是轻而易举的．从某种程度上来说，对简单的初值问题 $y' = y$, $y(0) = 1$ 也是如此．为了求得解的数值，人们必须计算或者查表，可能还要对 e^x 进行插值．另一个例子，方程 $y' = 1 - 2xy$ 的解为 $y(x) = e^{-x^2}\int_0^x e^{t^2}dt$．为了确定 $y(x)$ 的数值，人们必须计算一个积分，而它又不能用初等函数来表示，并且也没有合适的表可查．面对显式解的明显的局限性，数学家在用解析方法处理微分方程问题的早期，就开始使用具有更广泛适应性的近似方法．级数展开方法与 Picard-Lindeöf 迭代法是两个具有历史意义的例子．

微分方程的级数解的研究曾吸引了一些第一流的数学家，从而导致特殊解析函数理论的重要发展．目前，微分方程在什么情况下可以有简单的级数解，已经了如指掌．它们也不过形成了十分有限的一类．例如，虽然微分方程

$$y'' + \frac{A}{x}y' + \left(\frac{B}{x^2} + C + Dx^2\right)y = 0$$

(A, B, C, D — 常数)可以漂亮地用某种超越几何级数来解出，若将项 Dx^2 换成 Dx^3，则不存在这样简单的解．即使令级数解存在，对自变量的很多数值来说，它都可能收敛得很慢．在非线性微分方程的情形下，除了开头少数几项而外，要决定所有的系数，通常都是困难的．对于具有跳跃间断的方程来说，情况则更坏一些．

逐次逼近法是(偏和常)微分方程理论研究中的一个重要

工具，然而它作为解的数值计算的有效方法，其价值经常都是夸大其词的。作者没有见过一个微分方程可用逐次逼近法求其数值解而却不能以某些其它方法更方便地求解的例子.

0.3. 离散变量方法

本书将处理基于离散化原理的求常微分方程近似解的那些方法.这些方法的共同特点是，并不试图在自变量的整个连续区间上去逼近精确解 $y(x)$，只是在离散点 x_0，x_1，x_2，$\cdots$的一个集合上来考虑近似值. 通常点 x_n 不一定都是等距的.如果它们是等距的，我们就把它写成 $x_n = a + nh$，$n = 0, 1, 2, \cdots$. 量 h 称为步长大小，步长宽度或简称方法的步长. 一般来说，求解微分方程的离散变量法是由一个算法组成的，这个算法使每个网格点 x_n 对应一个数 y_n，而 y_n 被认为是在 x_n 点精确解 $y(x_n)$ 的近似值. 我们将考虑的主要问题之一是被称为离散误差 $e_n = y_n - y(x_n)$ 的这个量的大小，特别是它可作为步长 h 的一个函数.

与早期提到的一些方法相反，离散变量方法具有几乎是普遍可应用的优点. 例如，就初值问题 (0-4) 而言，对大多数离散变量方法的可应用性的唯一要求是对给定的 x 及 y 能计算出 $f(x,y)$ 的一个好的近似值[1]. 确实，为了保持离散误差充分小，可能需要对函数 $f(x, y)$ 进行多次的计算. 一度曾限制了离散变量方法的应用，今天就无需考虑这一点了. 当大量的数值计算，特别当计算具有多次重复性质时，可以在自动数字计算机上有效并可靠地得到完成.

在解初值问题的离散变量的方法中，我们可以将单步方法和多步方法加以区分. 在一个单步方法中，如果只知道

1) 例如，在某些情形下，所给出的函数 $f(x, y)$ 为经验的或是一个数值表，这一点就不是一个可有可无的要求了.

y_n就可以得到y_{n+1}，而无需知道或存储任何前面的值y_{n-1}，y_{n-2}，…．对于多步法来说，在计算y_{n+1}时需要知道的不仅是y_n，还要知道前面的一些值y_{n-1}，y_{n-2}，…．如果为了计算y_{n+1}，而要求知道$y_n, y_{n-1}, \cdots, y_{n+1-k}$值，则是$k$步方法．

在单步方法中，出发点没有特殊的作用．事实上，每一个网格点可以看作一个新的出发点，这就易于改变步长．多步方法要求一个特殊的初始处理，由于在点x_0，$x_1, \cdots, x_{k-1}$上对计算来说缺少必须具备的某些出发值y_{n-j}，同时在步长改变的点上也需要一个特殊的计算，所有这些便增加了所需机器代码的长度和复杂性．另一方面，多步方法可能较单步方法更为精确．这二类方法的进一步的相对优越性将在以后讨论．由于在这二种情形中基础理论的不同，故有必要分别处理这二类方法．第一章至第四章中考虑单步方法，第五章至第七章中考虑多步方法．

注

§0.2．在大多数微分方程书中都论述了级数展开法和逐次逼近法，例如 Levy 和 Baggott [1934]，Milne [1953]，Warga [1953]，Cernysenko [1958]，Franklin [1952] 都曾讨论了它们在数值解方面的应用．

第I部分 初值问题的单步方法

第一章 一阶单个方程的 Euler 方法

对于初值问题的近似解，Euler 方法（见 Euler [1913], p. 422; Euler [1914], p. 271）不仅在所有单步法中而且在所有方法中都是最简单的. 对于实际的数值问题，并不推荐 Euler 方法，因为它的精确度是十分有限的. 尽管如此，我们还将详细研究 Euler 方法，因为它能很明显的显示出对更为复杂的方法也具有的一些特性. 特别，这里介绍的离散误差和舍入误差的基本现象易于研究，因为这个方法分析简便. 全章都假设 h 为定步长.

1.1. 引言

1.1-1. Euler 方法的定义. 在 Euler 方法中，值 y_n 是按照如下公式：

$$y_0 = y(a) = \eta, \tag{1-1a}$$

$$y_{n+1} = y_n + hf(x_n, y_n), n = 0, 1, 2, \cdots \tag{1-1b}$$

递推计算出来的. 这些公式本身就给出一个明显的几何解释. 我们把微分方程 $y' = f(x, y)$ 看成在 (x, y) 平面的带形 $a \leqslant x \leqslant b$ 内确定方向场的一个方程. 于是求解微分方程的问题在几何上相当于确定这样一条曲线，它通过给定初始点

(x_0, y_0)，并且在其上每一点的斜率与由方向场所规定的斜率一致．由 (1-1) 确定的点 (x_n, y_n) 可作为折线的顶点，折线通过准确的初始点且具有这样的性质，每段都具有在它的左端点的方向场规定的方向．

公式 (1-1) 还可以有几个解析解释．

(i) 如果我们用向前差商在点 (x_n, y_n) 处逼近微分方程中的导数，则得到

$$\frac{y_{n+1} - y_n}{h} = f(x_n, y_n),$$

解出 y_n 便得 (1-1b)．

(ii) 在积分限 x 和 $x + k$ 之间积分恒等式

$$y'(t) = f(t, y(t)),$$

得到
$$y(x + k) - y(x) = \int_x^{x+k} f(t, y(t))\,dt.$$

特别，如果 $x = x_n$ 和 $k = h$，则

$$y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} f(t, y(t))\,dt.$$

用数值积分的粗糙公式（积分区间的长度乘被积函数在左端点上的值）来近似这个积分，并且把 $y(x_n)$ 恒同于 y_n，我们仍得(1-1)．

(iii) 假设在点 x_n 附近能够把解用 Taylor 级数展开：

$$y(x_n + h) = y(x_n) + hf(x_n, y(x_n)) + \frac{1}{2} h^2 y''(x_n) + \cdots.$$

公式 (1-1) 是截去这个级数在 h 的线性项以后的结果．

上述每一个解释均指出，在以后各章中要讨论的 Euler 方法一类推广途径．有趣的是，用(i)(数值微分) 所表示的推广，似乎是最直接的，但已经证明是三种推广中成效最少的．

1.1-2. 计算上的研究． Euler 方法的价值并非因为它作为一个计算过程，而是为了便于与以后所要讨论的更为复杂

的方法相比较．在图 1.1 中，我们用通常的表示法给出数字计算机上使用 Euler 方法求解初值问题 (0-4) 的框图．

由于同样的原因，我们在表 1.1 中列出用步长 $h=0.1$ 时 Euler 方法对初值问题

$$y' = x - y^2,\ y(0) = 0 \tag{1-2}$$

进行数值积分时的前几个值．在 §2.1-2 和 §5.1-2 中要用更精确的方法来处理这个同样的问题．箭头表示计算数值的顺序．

表 1.1　Euler 方法的数值例证

x_n	y_n		$f(x_n, y_n)$
0	0	→	0
		↙	
0.1	0	→	0.10000
		↙	
0.2	0.01000	→	0.19990
		↙	
0.3	0.02999	→	0.29910
		↙	
0.4	0.05990		

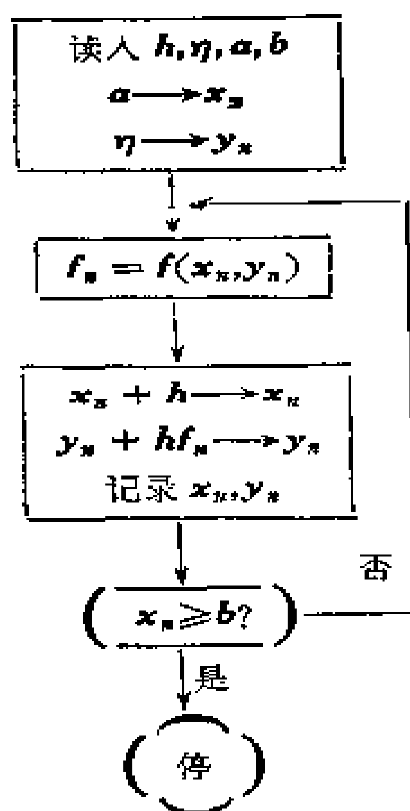


图 1.1　Euler 方法的框图

在求解微分方程的任何一个数值方法中，数值工作大部分用于对自变元的不同值计算出函数 $f(x, y)$ 的值．因此，通常每积分一步计算函数 $f(x, y)$ 值的次数便成为衡量任何一个方法所要求的计算工作量的标准．我们将称这个次数为置换的特定的数．对于 Euler 方法，这个次数显然为 1．

1.1-3. 一个解析例子．如果 $f(x, y)$ 是十分简单的函数，便有可能对 y_n 求解递推关系式 (1-1)，并且可求出 y_n 作为 n 及 h 的函数的显式表达式．这样一个显式解很少有实际意

义，因为它通常只在微分方程本身用封闭形式可求解的情形下才能求出来．但是，在研究所考察的方法的理论性质时，它是有帮助的．

我们来求初值问题 $y' = y$，$y(0) = 1$ 解的 Euler 近似显式形式．在这里有 $f(x,y) = y$，因此 (1-1b) 化成

$$y_{n+1} = y_n + hy_n = (1 + h)y_n, \quad n = 0,1,2,\cdots.$$

由于 $y_0 = 1$，我们求得 $y_1 = 1 + h$，$y_2 = (1 + h)y_1 = (1+h)^2$，一般地，

$$y_n = (1 + h)^n, \quad n = 1,2,\cdots.$$

因为 $n = x/h$，从而在点 $x_n = x$ 近似解的值由公式

$$y_n = (1 + h)^{x/h} = [(1 + h)^{1/h}]^x$$

给出．根据微积分中［Taylor［1955］，74 页］众所周知的定理，当 $h \to 0$ 时，它趋向于 e^x． 于是我们已经证明，在所考虑的特例中，当减小网格的大小时，便能任意好地逼近初值问题的精确解．

1.1-4．*数值方法的误差*． 求解微分方程的数值的逐步方法的误差应归结于两个来源． 第一个来源是由于用固定步长 h 的方法提供的数 y_n，即使计算到无穷多位小数，它与真解对应的值 $y(x_n)$ 是极少一致的． 这个差异必须预料到，因为一个数值算法事实上必定是有限次运算．即使是很简单的微分方程(例如 $y' = y$) 的解都是超越函数，因此不能用有限个有理运算来计算．差

$$e_n = y_n - y(x_n)$$

称为离散误差，通常也称为截断误差，这里 y_n 表示由算法给出的准确数值．

误差的第二个来源是由于在大多数情形下，数 y_n 不能计算到无限精确，因为任何计算工具都具有有限的精确度．我们记 $\tilde{y}_n$ 为代替 y_n 的真实计算的值．差

$$r_n = \tilde{y}_n - y_n$$

称为舍入误差.

利用三角不等式,我们求得总误差

$$|\tilde{y}_n - y(x_n)| = |(\tilde{y}_n - y_n) + (y_n - y(x_n))| \leqslant |e_n| + |r_n|.$$

为了区分 $\tilde{y}_n$ 和 y_n, 我们常常称 y_n 为对微分方程的真解的理论近似值, 而称 $\tilde{y}_n$ 为其数值近似值. 真解总是记成 $y(x)$.

一旦求解微分方程的方法在数学上完全确定时[用公式, 例如 (1-1)], 那么值 y_n 和离散误差 e_n 则可认为是完全确定的数. 另一方面,舍入误差 r_n 却不能由方法的数学公式来确定. 它们依赖于数字位数、机器所采用的数的系统、小数点的位置(定点或浮点运算)、数值运算所安排的次序、计算 $f(x,y)$ 所使用的子程序的精确度以及其它因素. 因此离散误差比舍入误差更容易适用于数学分析,这是不足为奇的. 但是,认为舍入误差永远是不可预料的,那是荒谬的,关于它的相当现实和实际的描述是可能达到的.

1.1-5. *各种类型的误差估计.* 求解数值问题(不仅对求解微分方程的问题) 的许多算法可以看成依赖于一个参数 p 的数值运算. 基本思想是取参数的极限值,例如说,当 $p\to\infty$, 便得到这个问题的"真解", 参数 p 可假设是连续或离散值. 在求解形如 $f(x) = 0$ 的方程的 Newton 方法中, 在迭代格式 $x_{p+1} = x_p - f(x_p)/f'(x_p)$ 中 x 的下标起着 p 的作用. 在固定区间上求解微分方程的问题中, p 以形式 $(b - a)/h$ 出现,其中 h 是所用方法的步长. 只有关于方法误差 $e(p)$ 的性态作为 p 的函数为已知, 数值方法才认为是满意的. 在误差研究中有若干标准是可能达到的.

(i) 可以简便地知道方法是收敛的[1]，意指

$$e(p) \to 0 \quad (p \to \infty).$$

(ii) 可以知道收敛速度的某些情形．比如说，它可以用某一个函数 $\varphi(p)$ 来表示，当 $p \to \infty$ 时，$\varphi(p)$ 趋向零，并且对一切充分大的 p 有一个未定常数 C，使得估计

$$|e(p)| \leqslant C\varphi(p)$$

成立．这个结果常常写成形式

$$e(p) = O(\varphi(p)).$$

读作：$e(p)$ 具有 $\varphi(p)$ 的阶．

(iii) 更为精确地说，需要知道

$$|e(p)| \leqslant \varphi(p) \quad 对一切 p \geqslant p_0,$$

其中 p_0 是指定的数，$\varphi(p)$ 仍是当 $p \to \infty$ 时而趋于零的一个指定的函数．这样的结果称为误差的界．

(iv) 最后我们知道：当 $p \to \infty$ 时

$$\frac{e(p)}{\varphi(p)} \to 1 \quad (p \to \infty),$$

其中仍有 $\varphi(p) \to 0$．这个公式称为误差的渐近公式．

上面陈述的几个结果，误差的界仅能使我们对所取的 p 值，使之保证误差 $|e(p)|$ 小于一个预先指定的数．对于包含微分方程在内的离散变量方法的数值分析的许多算法，都有可能得到误差的界．从实用观点出发，这些界常常是很少有用的，因为由函数 $\varphi(p)$ 给出的数值要比真正误差大得多．在这种情形便希望用渐近公式或者甚至用与误差的界有关的结果来补充这个界．在以后的章节中将有足够的机会来说明这一点．

---

1) 对于任何有用的方法都要求是收敛的，这几乎是不用说的了．但是，存在着一些数值上很有成效的近似方法（例如含大参数的定积分的渐近表达式），它并不满足收敛性的要求．

型 (ii), (iii), 和 (iv) 的结果可导出收敛性，并且 (i) 的结果其实际意义是很小的，当可得到其它类型的结果时。但是，型 (i) 的结果理论上的意义却是很大的。在许多情形，在比其它任何结果较弱的假设下，证明 (i) 是可能的。的确，有时无需假设问题的解的存在便可证明数值方法是收敛的。于是由数值方法的收敛性从而导出解的存在性。正如下面所要指出的，Euler 方法则是一个恰当的例子。

1.2. 初值问题解的存在性

在本节中，我们将给出基本初值问题 (0-4) 存在唯一解的初等的构造性的证明。

1.2. 定理的陈述． 我们假设实函数 $f(x, y)$ 满足以下条件：

(A) $f(x,y)$ 在带形 $a \leqslant x \leqslant b$，$-\infty < y < +\infty$ 内是确定且连续的，其中 a 与 b 都是有限数；

(B) 存在一个常数 L，使得对于任何 $x \in [a, b]$ 以及任何二个数 y 和 y^*，都有

$$|f(x,y) - f(x,y^*)| \leqslant L|y - y^*|. \qquad (1-3)$$

我们将证明：

定理 1.1. 令 $f(x,y)$ 满足上面的条件 (A) 和 (B)，且令 η 是一个给定的数，那么恰好存在一个具有以下三个性质的函数 $y(x)$：

(i) 对于 $x \in [a,b]$，$y(x)$是连续且可微的；

(ii) $y'(x) = f(x,y(x))$，$x \in [a,b]$；

(iii) $y(a) = \eta$.

简言之，初值问题 (0-4) 有唯一解。在 1.2-2 至 1.2-7 中将要给出的证明就是要证明用 Euler 方法确定的近似解的某

个序列收敛于函数 $y(x)$，而它具有所要求的性质.

上述条件 (B) 称为 Lipschitz 条件. 如果 $f(x, y)$ 在所考虑的带形内对 y 有连续导数且有界，则必然满足这个条件. 因为在这种情形，利用中值定理，我们便有

$$f(x,y)-f(x,y^*)=\frac{\partial f}{\partial y}(x,\bar{y})(y-y^*),$$

其中 $\bar{y}$ 是在 y 与 y^* 之间的一个值. 于是立即得到 (B). 但是，正如例 $f(x,y)=|y|$ 所指出，$\frac{\partial f}{\partial y}$ 的存在对 (B) 不是必要的. 另一方面，例 $f(x, y)=|y|^{\frac{1}{2}}$ 说明 f 的连续性不是充分的.

还要注意到，关于 (B) 的右端项 $L\cdot|y-y^*|$ 不能用形式为 $L|y-y^*|^\alpha (\alpha\neq 1)$ 的任何表达式来代替. 如果 $\alpha>1$，那么在整个区间 $[a,b]$ 内解就不一定存在. 正如问题 $(\varepsilon>0)$

$$y'=|y|^{1+\varepsilon},\ y(0)=1$$

所指出，其解 $y(x)=(1-\varepsilon x)^{-1/\varepsilon}$ 在 $x=\varepsilon^{-1}$ 就不再存在. 如果 $\alpha<1$，解不一定是唯一的，这可用初值问题

$$y'=|y|^{1-\varepsilon},\ y(0)=0$$
$$(0<\varepsilon<1)$$

来阐明. 除去解 $y(x)=0$ 外，它还有无穷多个解 $y(x)=0$，$0\leqslant x\leqslant c$；$y(x)=[\varepsilon(x-c)]^{1/\varepsilon}$，$x\geqslant c$，其中 $c>0$ 是任意的.

1.2-2. 近似解 $y_p(x)$ 的定义. 我们令

$$h_p=\frac{b-a}{2^p}\quad(p=0,1,2,\cdots),\tag{1-4}$$

并且用 $y_p(x)$ 表示逐段线性函数，其图形是由步长 $h=h_p$ 的 Euler 方法提供的近似解点 (x_n, y_n) 作为顶点的折线. 从而

每个相继的 $y_p(x)$ 可用步长折半的 Euler 方法来得到（见图 1.2）.

为了解析地描述函数 $y_p(x)$，对于任何 $x\in[a,b]$，我们用[1] $x_{[p]}$ 表示第 p 个划分中在 x 的左边最近的节点(但不与 x 重合)，记为

$$x_{[p]}=a+nh_p, \tag{1-5}$$

其中 n 是满足

$$a+nh_p<x\leqslant a+(n+1)h_p \tag{1-6}$$

的唯一整数．为避免复杂起见，对一切 p，令 $a_{[p]}=a$．我们注意到，对 $x>a$

$$x_{[p]}<x\leqslant x_{[p]}+h_p, \tag{1-7}$$

以及对 $p<q$

$$x_{[p]}\leqslant x_{[q]}. \tag{1-8}$$

利用这些记号，上面引进的函数可用如下关系式来定义：

$$y_p(a)=\eta, \tag{1-9}$$

$$y_p(x)=y_p(x_{[p]})+(x-x_{[p]})f(x_{[p]},y_p(x_{[p]})),$$
$$a<x\leqslant b.$$

我们注意函数 $y_p(x)$ 在区间 $[a,b]$ 内是连续的．

作为实现定理 1.1 证明的第一步，我们将证明：

引理 1.1. 函数序列 $y_p(x)$ 当 $p\to\infty$ 时，对 $x\in[a,b]$ 是一致地收敛于一个连续函数 $y(x)$.

通过证明对 $[a,b]$ 中的每一个 x 值，$y_p(x)$ $(p=0,1,2,\cdots)$ 形成一个 Cauchy 序列来完成引理的证明．更确切地说，我们将证明，对于每一个 $\varepsilon>0$，存在一个正数 $P=P(\varepsilon)$，使得对一切 $x\in[a,b]$，当 $p>P$，$q>P$ 时，

$$|y_p(x)-y_q(x)|<\varepsilon. \tag{1-10}$$

---

1) 仅在本节使用这个记号．

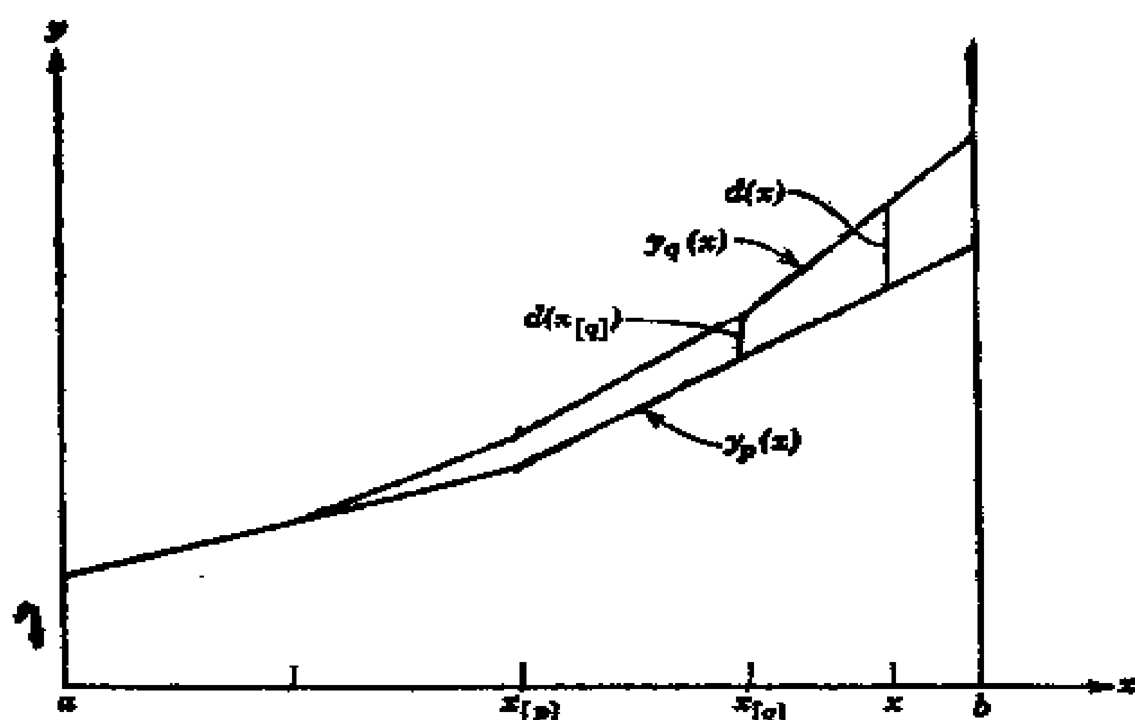


图 1.2 $p=1$，$q=2$ 时的近似解 $y_p(x)$ 和 $y_q(x)$

根据分析学中基本原理（见 Taylor [1955], p. 594，定理 1），于是存在一个函数 $y(x)$，使得 $\lim\limits_{p\to\infty} y_p(x) = y(x)$ 对 $x\in[a, b]$ 一致成立．利用另一个定理（见 Taylor [1955], p. 596，定理 II），连续函数序列的一致极限函数 $y(x)$ 本身是连续的．

在 1.2-5 中给出 (1-10) 的证明．在 1.2-3 和 1.2-4 中导出它所需要的某些辅助结果．

1.2-3. 递推不等式的解． 并非仅是为了引理 1.1 的证明，而且也为以后的需要，我们必须处理满足以下不等式的已知数列 $\xi_n (n=0,1,2,\cdots)$：

$$|\xi_{n+1}| \leqslant A|\xi_n| + B, \tag{1-11}$$

其中 A 和 B 都是与 n 无关的某一个非负常数．于是我们希望有一个对 $|\xi_n|$ 的估计，它是用 $|\xi_0|$ 而不是用 $|\xi_{n-1}|$ 来表示的．这样一个估计由以下引理给出：

辅助引理 1.2. 如果对于 $n=0, 1, 2, \cdots$，数 ξ_n 满足

(1-11)，那么

$$|\xi_n| \leqslant A^n|\xi_0| + \begin{cases} \dfrac{A^n - 1}{A - 1} B, & A \neq 1, \\ nB, & A = 1, \end{cases} \qquad (1\text{-}12)$$

$$(n=0,1,2,\cdots,N).$$

对于 $n = 1$，(1-12) 恒同于 (1-11)，从而由假设它是成立的．假设对于 $n < N$ 的值，(1-12) 成立，于是应用 (1-11)，我们求得(如果 $A \neq 1$)

$$\begin{aligned} |\xi_{n+1}| &\leqslant A\left\{A^n|\xi_0| + \frac{A^n - 1}{A - 1} B\right\} + B \\ &= A^{n+1}|\xi_0| + \left(A\,\frac{A^n - 1}{A - 1} + 1\right) B \\ &= A^{n+1}|\xi_0| + \frac{A^{n+1} - 1}{A - 1} B, \end{aligned}$$

这就是把 (1.12) 中的 n 增加 1． 如果 $A = 1$，可用类似的方法来证明这个结果．因此，由归纳法推出引理 1.2 成立．

在引理 1.2 的许多应用中，A 是取形式为 $1 + \delta$ 的量，其中 δ 是一个小的正数．于是我们可使用不等式

$$A = 1 + \delta < e^{\delta} \quad (\delta > 0),$$

它可由 e^{δ} 的 Taylor 展式直接推出，并把 (1-12) 写成形式

$$|\xi_n| \leqslant e^{n\delta}|\xi_0| + \frac{e^{n\delta} - 1}{\delta} B. \qquad (1\text{-}13)$$

这种形式有时使用起来要比 (1-12) 更为方便．

为了得到用任意步长 h 对初值问题 (0-4) 的解所得到的值 y_n 的界，我们应用引理 1.2． 先从所提到的条件 (B) 来着手，如果取 $y^* = 0$，就有

$$|f(x,y) - f(x,0)| \leqslant L|y|.$$

从而得到 $|f(x,y)| \leqslant L|y| + c$，其中

$$c = \max_{x\in[a,b]} |f(x,0)|. \qquad (1\text{-}14)$$

从递推关系式(1-1b)，我们容易求得

$$|y_{n+1}| \leqslant |y_n| + h|f(x_n, y_n)|$$
$$\leqslant (1 + hL)|y_n| + hc, \quad n = 0, 1, 2, \cdots.$$

利用 $|y_0| = |\eta|$，并应用形式为(1-13)的引理 1.2 的结果，使 y_n 等同于 ξ_n，δ 等同于 hL 以及 B 等同于 hc，我们得到

$$|y_n| \leqslant e^{nhL}|\eta| + \frac{e^{nhL} - 1}{L} c.$$

由于 $0 \leqslant nh = x_n - a \leqslant b - a$，我们求得估计

$$|y_n| \leqslant Y, \quad x_n \in [a, b], \tag{1-15}$$

其中常数

$$Y = e^{(b-a)L}|\eta| + \frac{e^{(b-a)L} - 1}{L} c \tag{1-16}$$

与 h 无关．由于函数 $y_p(x)(p = 0, 1, 2, \cdots)$ 是由某一个特殊的 h 值得到的点 (x_n, y_n) 连成直线段组成的，我们有

$$|y_p(x)| \leqslant Y, \quad x \in [a, b], \tag{1-17}$$

Y 用相同的定义．

最后结论的意义在于，它容许我们陈述这样的事实，即近似解 $y_p(x)$ 全部在 (x, y)-平面的某一个紧致的(即有界和闭的)区域内．

1.2-4. 连续性模． 我们用 R 表示在 (x, y)-平面内不等式 $a \leqslant x \leqslant b$，$|y| \leqslant Y$ 所确定的矩形区域，其中 Y 由(1-16)确定．由假设(A)，$f(x, y)$ 在 R 内是连续的．对于任何 $\delta > 0$，现在我们规定量

$$\omega(\delta) = \max|f(x, y) - f(x^*, y)|, \tag{1-18}$$

其中最大值是对于使得 $|x - x^*| \leqslant \delta$ 的 R 内的一切点 (x, y) 和 (x^*, y) 来取的．我们称 $\omega(\delta)$ 为函数 $f(x, y)$ 的连续性模[1]．

---

1) 更确切地说，称它为一致连续模，"连续性模"这个术语通常是仅当 f 为单个变量函数时才使用．

由这个定义，立即推得

(i) 如果 $0 \leqslant \delta \leqslant \delta'$，那么

$$\omega(\delta) \leqslant \omega(\delta'). \tag{1-19}$$

还要求函数 $\omega(\delta)$ 具有如下的简单性质.

(ii) 如果 $\alpha > 0$，$\beta > 0$，那么

$$\omega(\alpha+\beta) \leqslant \omega(\alpha) + \omega(\beta). \tag{1-20}$$

证. 在下面的所有表达式中，最大值都是对于 R 内的一切点 (x, y)，(x^{**}, y) 和 (x^*, y) 来取的，这样取是受到所指出的限制的. 我们有

$$\begin{aligned}
\omega(\alpha+\beta) &= \max_{|x-x^*|\leqslant\alpha+\beta} |f(x,y)-f(x^*,y)| \\
&\leqslant \max_{\substack{|x-x^{**}|\leqslant\alpha \\ |x^{**}-x^*|\leqslant\beta}} \{|f(x,y)-f(x^{**},y)| + |f(x^{**},y)-f(x^*,y)|\} \\
&\leqslant \max_{|x-x^{**}|\leqslant\alpha} |f(x,y)-f(x^{**},y)| \\
&\quad + \max_{|x^{**}-x^*|\leqslant\beta} |f(x^{**},y)-f(x^*,y)| \\
&= \omega(\alpha)+\omega(\beta),
\end{aligned}$$

于是证明了(1.20).

(iii) $$\lim_{\delta\to 0}\omega(\delta) = 0.$$

证. 如果结论不成立，就存在一个数 $\theta > 0$ 以及完全属于 R 内的二个点列 (x_n, y_n) 与 (x_n^*, y_n^*)，使得对于

$$n = 1, 2, \cdots,$$

都有

$$|x_n - x_n^*| < 1/n \tag{1-21}$$

及

$$|f(x_n, y_n) - f(x_n^*, y_n)| \geqslant \theta. \tag{1-22}$$

由于 R 是紧致的，根据 Bolzano-Weierstrass 定理（见 Taylor [1955], p. 487)，序列 (x_n, y_n) 在 R 内有一个极限点 (x, y)，并且我们可取出一个子序列，其元素记成 (x_{n_k}, y_{n_k}) $(k = 1,$

2, …), 它收敛于(x, y). 由(1-22), 对每一个下标k, $f(x_{n_k}, y_{n_k})$和$f(x_{n_k}^*, y_{n_k})$中至少有一个值与$f(x,y)$相差最少为$\theta/2$,这就与f在点(x, y)的连续性不相容. 由这个矛盾便证明了(iii).

(iv) 函数$\omega(\delta)$对于$\delta \geqslant 0$是连续的.

证. 对于$\delta = 0$, 连续性由(iii)直接推出. 对于$\delta > 0$, 为了证明连续性, 令$\varepsilon > 0$是给定的. 由(iii), 存在$\theta > 0$, 使得对于$0 \leqslant \gamma < \theta$, 有$\omega(\gamma) < \varepsilon$. 利用(i)与(ii), 从而有

$$0 \leqslant \omega(\delta + \gamma) - \omega(\delta) < \varepsilon,$$

并且对于$0 \leqslant \gamma < \theta$, 同样有

$$0 \leqslant \omega(\delta) - \omega(\delta - \gamma) = \omega(\delta - \gamma + \gamma) - \omega(\delta - \gamma) < \varepsilon.$$

综合这些关系式,推得

$$|\omega(\delta + \gamma) - \omega(\delta)| < \varepsilon, \quad |\gamma| < \theta.$$

于是证明了在点δ是连续的.

1.2-5. 引理 1.1 的证明. 现在我们对任意$x \in [a, b]$来估计差

$$d(x) = y_q(x) - y_p(x) \qquad (1\text{-}23)$$

(见图 1.2), 其中p与q是二个非负整数, $p < q$. 这个估计是通过对$d(x_{[q]})$的估计来得到的. 以下引理对这两个估计是有帮助的.

引理 1.3. 对于$t \in [a, b]$

$$|d(t)| \leqslant [1 + (t - t_{[q]})L]\,|d(t_{[q]})| + (t - t_{[q]})\Omega_p, \qquad (1\text{-}24)$$

其中

$$\Omega_p = \omega(h_p) + LMh_p, \quad M = \max_{(x,y)\in R} |f(x,y)|. \qquad (1\text{-}25)$$

证. 我们注意到，利用 (1-9)，就有

$$y_q(t) - y_q(t_{[q]}) = (t - t_{[q]})f(t_{[q]}, y_q(t_{[q]})),$$
$$y_p(t) - y_p(t_{[q]}) = (t - t_{[q]})f(t_{[p]}, y_p(t_{[p]})).$$

从第一个方程减去第二个方程，我们求得

$$d(t) - d(t_{[q]}) = (t - t_{[q]})[f(t_{[q]}, y_q(t_{[q]})) - f(t_{[p]}, y_p(t_{[p]}))].$$

在括号内的表达式可写成形式

$$f(t_{[q]}, y_q(t_{[q]})) - f(t_{[q]}, y_p(t_{[q]})) + f(t_{[q]}, y_p(t_{[q]})) - f(t_{[p]}, y_p(t_{[q]})) + f(t_{[p]}, y_p(t_{[q]})) - f(t_{[p]}, y_p(t_{[p]})),$$

这里每一行中的差可以单独估计. 由条件 (B)，第一个差是以 $L\ |d(t_{[q]})|$ 为界. 第二个差以 $\omega(\delta)$ 为界，其中

$$\delta = t_{[q]} - t_{[p]}.$$

利用 (1-7) 和 (1-19)，从而它以 $\omega(h_p)$ 为界. 由 (B)，第三个差的绝对值不超过

$$L\,|y_p(t_{[q]}) - y_p(t_{[p]})|,$$

并且由 (1-9)，这个数是以 LMh_p 为界. 于是立即推出引理 1.3 的结果.

我们利用引理 1.3，取 $t = t_\nu = a + \nu h_q$，$\nu = 1, 2, \cdots, k$，其中 $t_k = x_{[q]}$. 在这种情形，$t - t_{[q]} = h_q$，并且由(1-24)推得

$$|d(t_\nu)| \leqslant A\,|d(t_{\nu-1})| + B,$$

其中 $A = 1 + h_q L$，$B = h_q \Omega_p$. 应用辅助引理 1.2，取 $\xi_\nu = d(t_\nu)$. 由于 $d(a) = 0$，我们得到

$$|d(x_{[q]})| \leqslant \Omega_p \frac{e^{(x_{[q]}-a)L} - 1}{L}. \tag{1-26}$$

利用引理 1.3，取 $t = x$，求得

$$|d(x)| \leqslant [1 + (x - x_{[q]})L]\,|d(x_{[q]})| + (x - x_{[q]})\Omega_p.$$

由于

$$1+(x-x_{[q]})L \leqslant e^{(x-x_{[q]})L},$$

最后不等式可以放宽为

$$|d(x)| \leqslant e^{(x-x_{[q]})L}|d(x_{[q]})| + \frac{e^{(x-x_{[q]})L}-1}{L}\Omega_p,$$

把它与 $d(x_{[q]})$ 的估计 (1-26) 结合起来，得

$$|d(x)| \leqslant \Omega_p\left\{e^{(x-x_{[q]})L}\frac{e^{(x_{[q]}-a)L}-1}{L}+\frac{e^{(x-x_{[q]})L}-1}{L}\right\},$$

或者，化简成

$$|d(x)| \leqslant \Omega_p\frac{e^{(x-a)L}-1}{L}. \qquad (1-27)$$

这便是关于 $|d(x)|$ 所需要的估计.

(1-27) 的右端的表达式不依赖于 q. 由于利用 1.2-4 的结论 (iii)，当 $p\to\infty$ 时，$\Omega_p=\omega(h_p)+h_pLM\to 0$，故取 p 充分大时，Ω_p 就可以任意地小. 从而满足 Cauchy 准则 (1-10)，并证明了引理 1.1.

1.2-6. 证明 $y(x)$ 是一个解. 现在我们再来证明极限函数 $y(x)$ 是可微的且满足给定的微分方程. 在没有使用积分概念这个意义上来说，我们的证明则完全是初等的.

我们需要一些函数 $f_p(x)$，它对 $p=0,1,2,\cdots$ 定义为

$$f_p(x)=\begin{cases} f(a,\eta), & x=a,\\ f(x_{[p]},y_p(x_{[p]})), & a<x\leqslant b.\end{cases}$$

函数 $f_p(x)$ 在每一个区间 $x_{[p]}<x\leqslant x_{[p]}+h_p$ 内皆为常数，并且它们的常数值是近似解 $y_p(x)$ 在这个区间内的斜率. 我们断定

$$\lim_{p\to\infty}f_p(t)=f(t,y(t)) \quad 对\ a\leqslant t\leqslant b\ 一致地成立. \qquad (1-28)$$

事实上

$$f_p(t) - f(t, y(t)) = f(t_{[p]}, y_p(t_{[p]})) - f(t_{[p]}, y_p(t))$$
$$+ f(t_{[p]}, y_p(t)) - f(t_{[p]}, y(t))$$
$$+ f(t_{[p]}, y(t)) - f(t, y(t)),$$

在右端的三个差分别是以量 LMh_p，

$$[\omega(h_p) + LMh_p][e^{(b-a)L} - 1]$$

以及 $\omega(h_p)$ 为界，它们不依赖于 t 并且通过选取足够大的 p，可以使其任意地小。

现在我们将得到对量

$$\frac{y(z) - y(x)}{z - x} - f(x, y(x))$$

的估计，其中 $a \leqslant x < z \leqslant b$。令

$$\Delta_p = z_{[p]} - x_{[p]}, \quad \nu_p = \Delta_p / h_p,$$

利用 (1-9)，对于 $\nu_p > 1$，有

$$y_p(z_{[p]}) - y_p(x_{[p]}) = \sum_{\nu=1}^{\nu_p} \{y_p(x_{[p]} + \nu h_p)$$
$$- y_p(x_{[p]} + (\nu - 1)h_p)\}$$
$$= h_p \sum_{\nu=0}^{\nu_p - 1} f_p(x_{[p]} + \nu h_p)$$
$$= h_p \sum_{\nu=0}^{\nu_p - 1} \{f_p(x_{[p]} + \nu h_p) - f_p(x_{[p]})\}$$
$$+ \Delta_p f_p(x_{[p]}).$$

利用估计式

$$|f_p(s) - f_p(t)| \leqslant |f(s_{[p]}, y_p(s_{[p]})) - f(t_{[p]}, y_p(s_{[p]}))|$$
$$+ |f(t_{[p]}, y_p(s_{[p]})) - f(t_{[p]}, y_p(t_{[p]}))|$$
$$\leqslant \omega(|s_{[p]} - t_{[p]}|) + LM|s_{[p]} - t_{[p]}|,$$

求得

$$|f_p(x_{[p]} + \nu h_p) - f_p(x_{[p]})| \leqslant \omega(\Delta_p) + LM\nu h_p.$$

于是，利用 $\sum_{\nu=0}^{\nu_p-1}\nu\leqslant\frac{1}{2}\nu_p^2-\frac{1}{2}h_p^{-2}\Delta_p^2$，

$$
\begin{aligned}
|y_p(z_{[p]})-y_p(x_{[p]})-\Delta_p f_p(x_{[p]})|
&\leqslant h_p\sum_{\nu=0}^{\nu_p-1}[\omega(\Delta_p)+LMh_p\nu]\\
&\leqslant \Delta_p\omega(\Delta_p)+\frac{1}{2}LM\Delta_p^2.
\end{aligned}
$$

令 $p\to\infty$，把 x 和 z 固定，有

$$x_{[p]}\to x,\quad z_{[p]}\to z,\ \text{从而}\ \Delta_p\to\Delta=z-x$$

而且

$$y_p(z_{[p]})\to y(z),\quad y_p(x_{[p]})\to y(x).$$

并且由于 (1-27)，

$$f_p(x_{[p]})=f_p(x)\to f(x,y(x)).$$

我们再利用 $\omega(\delta)$ 为连续(见 §1.2-4) 的事实，因此

$$|y(z)-y(x)-\Delta f(x,y(x))|\leqslant\Delta\omega(\Delta)+\frac{1}{2}LM\Delta^2,$$

或者，以 Δ 除之，

$$\left|\frac{y(z)-y(x)}{\Delta}-f(x,y(x))\right|\leqslant\omega(\Delta)+\frac{1}{2}LM\Delta.$$

由于右端的表达式可以通过选取充分小 Δ 变得任意地小，这就证明了 $y(x)$ 存在右导数并且取值 $f(x,y(x))$. 从 $f(x,y(x))$ 的连续性推出 $y(x)$ 也存在左导数且取相同的值.

$y(x)$ 满足微分方程的一个简短但却不是初等的证明，是建立在定积分的概念上(见 §3.1-5).

1.2-7. 证明 $y(x)$ 是唯一解. 上面证明了函数序列 $y_p(x)$ 收敛于初值问题 $y'=f(x,y)$，$y(a)=\eta$ 的某一个解 $y(x)$. 但同一个问题仍可能有其它解. 为了完成定理 1.1 的证明，我们指出：如果 $z(x)$ 是初值问题的任意一个解，序

列 $\{y_p(x)\}$ 必定收敛于 $z(x)$，这就证明了它不可能有其它的解.

根据解的定义，$z(x)$在 $[a,b]$ 上是连续的. 由 (A) 推得 $z'(x)=f(x,z(x))$ 在相同的闭区间上也是连续的. 令

$$Z_1=\max_{x\in[a,b]}|z'(x)|, \tag{1-29}$$

$$d(x)=y_p(x)-z(x),$$

并且采用十分类似于 1.2-5 中的方法来估计这个差. 对于任意 $t\in[a,b]$，有

$$y_p(t)-y_p(t_{[p]})=(t-t_{[p]})f(t_{[p]},y_p(t_{[p]})),$$

并利用中值定理，如果 τ 表示在 $t_{[p]}$ 与 t 之间的某一个数，便有

$$z(t)-z(t_{[p]})=(t-t_{[p]})f(\tau,z(\tau)).$$

从前一个关系式减去后一个关系式并且插入适当的值，求得

$$\begin{aligned}d(t)-d(t_{[p]})=(t-t_{[p]})\{&f(t_{[p]},y_p(t_{[p]}))-f(\tau,y_p(t_{[p]}))\\&+f(\tau,y_p(t_{[p]}))-f(\tau,z(t_{[p]}))\\&+f(\tau,z(t_{[p]}))-f(\tau,z(\tau))\},\end{aligned}$$

在括号内的三个差的界为 $\omega(h_p)$，$L|d(t_{[p]})|$ 以及通过再次应用 Lipschitz 条件和中值定理，界为 LZ_1h_p. 于是求得

$$\begin{aligned}|d(t)|\leqslant[1+(t-t_{[p]})L]|d(t_{[p]})|\\+(t-t_{[p]})\Omega_p',\end{aligned} \tag{1-30}$$

其中

$$\Omega_p'=\omega(h_p)+LZ_1h_p. \tag{1-31}$$

正如从 (1-24) 导出 (1-27) 那样，使用完全相同的方法，由 (1-30) 推得界

$$|d(x)|\leqslant\Omega_p'\frac{e^{(x-a)L}-1}{L}. \tag{1-32}$$

因为当 $p\to\infty$ 时，$\Omega_p'\to0$，从而 $y_p(x)\to z(x)$. 因此

$$z(x)=y(x).$$

由此便完成定理 1.1 的证明.

1.2-8. *存在定理的一些特殊情形.* 特别是，如果函数 $f(x,y)$ 不依赖于 y，并且在区间 $[a,b]$ 内是 x 的连续函数，那么条件 (A) 和 (B) 就被满足. 因此，我们已经证明了连续函数不定积分的存在性. 以上给出的证明确实是这个基本定理的有效证明，因为它从未用到积分概念.

满足条件的另一种情形是线性微分方程

$$y' = g(x)y + p(x),$$

其中 $g(x)$ 和 $p(x)$ 都是 $[a, b]$ 内的连续函数. 于是 Lipschitz 条件 (B) 成立，取

$$L = \max_{x \in [a,b]} |g(x)|.$$

在对于大的 $|y - y^*|$ 值违反条件 B 的情形下，例如

$$f(x,y) = x^2 + y^2,$$

有时仍然可用以下灵活的试验有可能断定解的存在性. 对于大的 y 值，则可改变 $f(x,y)$ 的定义，如令 $f(x, y) = f(x,Y)$ $(y \geq Y)$ 及 $f(x,y) = f(x,-Y)(y \leq -Y)$，其中 Y 是一个适当(大的)数. 从而便可应用存在性定理. 如果解 $y(x)$ 满足 $|y| \leq Y$，这就是原来给出的微分方程的解.

今后我们总是假设给定初值问题的解 $y(x)$ 是存在的. 如果对于不受限制的 y 值，Lipschitz 条件不成立，我们将假设它至少在解 $y(x)$ 的一个邻域内是成立的. 在这种情形，理论上的分析仅用于在那个邻域范围内.

1.3. Euler 方法的离散误差

1.3-1. *几个先验界.* 所谓近似的误差的先验界是意指，由问题的数据可以直接计算出的(原则上)界，而无需首先确

定近似解本身. 在目前情形,问题的数据是由函数 $f(x,y)$ 及由值 $y(a)$ 提供的 η 所组成的.

第一个先验界可以从定理 1.1 的证明中得到. 如果

$$e_n = y_n - y(x_n)$$

表示用步长 $h = h_p$ 得到的近似值 y_n 在点 $x = x_n = a + nh$ 上的误差,那么关系式 (1-32) 便导出

$$|e_n| \leqslant [\omega(h_p) + h_p L Z_1] \frac{e^{(x_n-a)L} - 1}{L}, \quad x_n \in [a,b],$$

这里 Z_1 表示 $|z'(x)|$ 的一个上界, 其中 $z(x)$ 为初值问题的任意一个解. 由于现在证明了唯一性,我们就有 $z(x) = y(x)$,它是折线函数 $y_p(x)$ 的极限. 每一个 $y_p(x)$ 保留在 1.2-4 中规定的矩形区间 R 内; 从而对于极限函数 $y(x)$ 同样也是成立的. 因为 $y'(x) = f(x,y(x))$, 所以常数 Z_1 可用 M 来代替,其中

$$M = \max_{(x,y)\in R} |f(x,y)|. \tag{1-33}$$

其次, 虽然 (1-32) 仅是就 $h = h_p$ 来证明的,实际上同样的证明对于任意 $h > 0$ 都适用. 如果表达式 $[e^{(x_n-a)L} - 1]L^{-1}$ 用 $x_n - a$ 来代替,那么对于为零的 Lipschitz 常数, 所得到的估计也是正确的. 由于常常出现指数表达式及当 $L \to 0$ 时它的极限形式, 所以方便的办法是, 对于任意的 $x \geqslant 0$ 及 $L \geqslant 0$, 定义函数

$$E_L(x) = \begin{cases} \dfrac{e^{Lx} - 1}{L}, & L > 0, \\ x, & L = 0. \end{cases}$$

这个函数将看成 Lipschitz 函数.

于是所得结果可陈述如下:

定理 1.2. 如果 $f(x, y)$ 满足定理 1.1 的条件, 那么用一个任意步长 $h > 0$, Euler 方法确定的近似值 y_n 的误差 e_n 满

足

$$|e_n| \leqslant [\omega(h) + hLM]E_L(x_n - a), \quad x_n \in [a,b], \tag{1-34}$$

其中 $\omega(h)$ 和M分别是由 (1-18) 及 (1-33) 确定的.

界 (1-34) 实际上是不容易应用的,因为出现了有些难于估计的函数 $\omega(h)$. 如果已知准确解有连续二阶导数,那么可得到更明确的结果.

定理 1.3. 令函数 $f(x,y)$ 满足条件 (B), 又令初值问题的准确解 $y(x)$ 在 $[a,b]$ 内是二次连续可微. 如果

$$N(x) = \frac{1}{2} \max_{t \in [a,x]} |y''(t)|, \tag{1-35}$$

那么 Euler 方法的误差满足

$$|e_n| \leqslant hN(x_n)E_L(x_n - a). \tag{1-36}$$

上述结果对于近似值仅取在节点上的更为精确的方法的类似定理来说是具有典型性的. 我们将要给出上述定理不涉及到存在定理的简单证明, 它对于更为复杂的情形来说也是典型的.

按定义

$$y_{m+1} = y_m + hf(x_m, y_m), \quad m = 0,1,2,\cdots n-1,$$

利用 Taylor 公式 (见 Taylor [1955], p. 112, 定理 III), 对于准确解,有

$$y(x_{m+1}) = y(x_m) + hf(x_m, y(x_m)) + \frac{1}{2}h^2y''(\xi),$$

其中 $x_m < \xi < x_{m+1}$.

相减后,求得误差 $e_m = y_m - y(x_m)$:

$$e_{m+1} = e_m + h[f(x_m, y_m) - f(x_m, y(x_m))] - \frac{1}{2}h^2y''(\xi).$$

取绝对值,便得到

$$|e_{m+1}| \leqslant |e_m| + h|f(x_m, y_m) - f(x_m, y(x_m))| + \frac{1}{2} h^2 |y''(\xi)|. \quad (1\text{-}37)$$

用条件 (B) 来估计右端的第二项，结果是

$$|f(x_m, y_m) - f(x_m, y(x_m))| \leqslant L|y_m - y(x_m)| = L|e_m|.$$

对于 $x_{m+1} \leqslant x_n$，用 $N(x_n)$ 来估计第三项，得到的 $|e_{m+1}|$ 的界可写成

$$|e_{m+1}| \leqslant A|e_m| + B, \quad m = 0, 1, 2, \cdots, n-1, \quad (1\text{-}38)$$

其中

$$A = 1 + hL, \quad B = h^2 N(x_n).$$

在这里要应用引理 1.2. 利用 $e_0 = 0$（无初始误差），得到

$$|e_n| \leqslant B \frac{A^n - 1}{A - 1} \leqslant hN(x_n) E_L(x_n - a),$$

这便是所需要的结果.

剩下的是估计 $N(x)$ 的问题. 因为近似的解是未知的，定义 (1-35) 当然不能使用. 但是，假设 $f(x,y)$ 在区域 R 内有连续的一阶偏导数，我们便可以微分恒等式

$$y'(x) = f(x, y(x)),$$

求得

$$\begin{aligned} y''(x) &= f_x(x, y(x)) + f_y(x, y(x)) y'(x) \\ &= f_x(x, y(x)) + f_y(x, y(x)) f(x, y(x)). \end{aligned}$$

因此

$$2N(x) \leqslant \max_{(x,y) \in R} |f_x + f_y f|.$$

我们用与定理 1.3 类似而在稍许放宽的条件下是正确的结果来结束这一节.

定理 1.4. 令 $y(x)$ 和 $N(x)$ 如定理 1.3 中所定义的，又令 $f(x,y)$ 满足相同条件. 令 $\{y_n\}$ 为满足

$$y_0 = \eta, \quad (1\text{-}39)$$

$$y_{n+1} = y_n + hf(x_n, y_n) + \theta_n h^2 C, \quad n = 0, 1, 2, \cdots$$

的任意数列，其中 $C \geqslant 0$ 为一常数，并且 θ_n 是每步都可以不同但总是满足 $|\theta_n| < 1$ 的数，那么

$$|y_n - y(x_n)| \leqslant h(N(x_n) + C)E_L(x_n - a). \qquad (1\text{-}40)$$

定理 1.4 的要点在于指出，即使递推关系式 (1-1) 不是准确地满足，如果 (1-1) 和 (1-39) 之间相差不很大，值 y_n 仍可收敛于 $y(x_n)$。暂且把量 $\theta_n h^2 C$ 看成舍入误差，虽然在以后定理 1.4 的应用中它们含意有所不同。

定理 1.4 的证明仍可使用与定理 1.3 的证明相同的方法。代替 (1-37)，我们得到

$$|e_{m+1}| \leqslant |e_m| + h|f(x_m, y_m) - f(x_m, y(x_m))| + \frac{1}{2}h^2|y''(\xi)| + h^2 C, \qquad (1\text{-}41)$$

这是因为在 (1-39) 的左端出现附加项。利用 Lipschitz 条件，我们把 (1-41) 写成形式 (1-38)，其中 A 如前，但 B 现在却为 $h^2 N(x_n) + C$。象上面一样来完成这个证明，我们便得 (1-40)。

1.3-2. 数值例子. 估计式 (1-36) 对许多经典的误差估计是很典型的. 在用步长 h 整除 $x - a$ 的固定点 x 上的误差表明，用依赖于 x 而与 h 无关的表达式乘以 h 的幂(这里为一次幂)为界. 但是，它绝不意味着，在任何具体情形下，误差象 (1-36) 所指出的那样大.

作为一个例子，我们考察在初值条件 $y(0) = 1$ 下，二个微分方程

$$y' = y, \qquad (1\text{-}42)$$

$$y' = -y \qquad (1\text{-}43)$$

的数值解. 显然真解为 $y = e^x$ 和 $y = e^{-x}$. 我们把真实误差与由 (1-36) 给出的界相比较. 对这二个方程，$L = 1$. 我

们对理论上的界是清楚的，并且可由真解确定出 N. 由于二阶导数分别为 e^x 和 e^{-x}, 我们有

对 (1-42),

$$2N(x) = e^x;$$

对 (1-43),

$$2N(x) = 1.$$

因此误差分别是以表达式

$$\frac{1}{2} he^x(e^x - 1), \qquad (1-44)$$

$$\frac{1}{2} h(e^x - 1) \qquad (1-45)$$

为界. 表 1.2a 和 1.2b 中对这些理论误差界与真正误差进行了比较. 在这两个计算中取了充分多的位数，使得在表中给出数字其结果都是准确的.

这些结果明确地指出，在所考虑的两个情形，界 (1-36) 并未给出真实的结果. 如果要确定 e^{-5} 使解 (1-43) 的误差不超过 10^{-3}, 那么界 (1-36) 便要求选取步长 h, 使得

$$\frac{1}{2} h(e^5 - 1) \leqslant 10^{-3},$$

即 $h \leqslant 1/73707$. 实际上, $h = 1/64$ 就足够了，如表 1.2b 所示.

在导出界 (1-45) 中，我们利用了给定方程的准确解的知识. 在更现实情形，其准确解并不知道，可以预料到理论上的

表 1.2a $h = 2^{-4}$, 用 Euler 方法对 $y' = y$, $y(0) = 1$ 的积分

x_n	1	2	3	4	5
y_n	2.69735	7.27567	19.62499	52.93537	142.7850
e_n	−0.02093	−0.11339	−0.46055	−1.66278	−5.6282
误差界(1-44)	0.03649	0.36882	2.99487	22.86218	170.9223

表 1.2b　$h = 2^{-6}$，用 Euler 方法对 $y' = -y$，$y(0) = 1$ 的积分

x_n	1	2	3	4	5
y_n	0.364987	0.133215	0.048622	0.017746	0.006477
e_n	−0.002892	−0.002120	−0.001165	−0.000570	−0.000261
误差界(1-45)	0.013424	0.049914	0.149106	0.418735	1.151666

界会给出更差的结果。因此很清楚地表明需要求得更现实的截断误差估计。

1.3-3. **误差的一个后验界.** 现在我们导出一个结果，它不是说明想得到 Euler 方法误差可能有多大，而是要指出误差真正有多大(近似地). 后面对更加精确的方法也给出类似的结果.

不等式 (1-36) 可以说明，当 $h \to 0$，$x_n = a + nh$ 固定时，数 $h^{-1}|e_n|$ 是有界的. 让我们来看看对特殊问题

$$y' = -y,\ y(0) = 1$$

的值 $h^{-1}e_n$. 表 1.3 给出对 $x_n = 1$ 取步长 $h = 2^{-k}$，$k = 1, 2, \cdots$ 计算出的 y_n，e_n 及 $h^{-1}e_n$ 的值.

从这个表看出，数 $h^{-1}e_n$ 不仅是有界的，而且还趋向一个

表 1.3

$k = -\log_2 h$	y_n	e_n	$h^{-1}e_n$
1	0.25000	−0.117879	−0.235758
2	0.316406	−0.051473	−0.205893
3	0.343509	−0.024270	−0.194164
4	0.356074	−0.011805	−0.188885
5	0.362055	−0.005824	−0.186373
6	0.364987	−0.002892	−0.185147
7	0.366438	−0.001441	−0.184540
8	0.367160	−0.000719	−0.184239
⋮	⋮		
∞	0.367879		

确定的极限. 我们应当指出，不仅对目前的特殊情形，而且用 Euler 方法对任何微分方程的近似解都确实如此. 在目前情形，这个极限将证明为 $-\frac{1}{2}e^{-1}=-0.183940$.

我们假设函数 $f(x,y)$ 不仅满足条件 (A) 和 (B)，而且在区域 R (见 1.2-4) 内还有连续的一阶和二阶导数. 在这些假设下，存在着真解 $y(x)$ 的三阶导数，并且可以写成

$$y(x_{n+1})=y(x_n)+hf(x_n,y(x_n))+\frac{1}{2}h^2y''(x_n)$$
$$+\frac{1}{6}h^3y'''(\xi),$$

其中 $x_n\leqslant\xi\leqslant x_{n+1}$. 从近似值所满足的对应关系式

$$y_{n+1}=y_n+hf(x_n,y_n)$$

减去这个关系式，我们得到[其中 $y_n=y(x_n)+e_n$]

$$e_{n+1}=e_n+h[f(x_n,y(x_n)+e_n)-f(x_n,y(x_n))]$$
$$-\frac{1}{2}h^2y''(x_n)-\frac{1}{6}h^3y'''(\xi). \qquad (1-46)$$

利用 Taylor 公式，括号内的表达式可写成形式

$$f_y(x_n,y(x_n))e_n+\frac{1}{2}f_{yy}(x_n,y^*)e_n^2,$$

其中 y^* 是 $y(x_n)$ 和 y_n 之间的一个值. 我们用 h 除以所得的关系式，并引入量

$$\bar{e}_n=h^{-1}e_n,$$

称它为伸缩误差. 利用前面的结果，伸缩误差是以常数

$$C_1=NE_L(b-a)$$

为界. 方程 (1-46) 可写成形式

$$\bar{e}_{n+1}=\bar{e}_n+h\left[f_y(x_n,y(x_n))\bar{e}_n-\frac{1}{2}y''(x_n)\right]+h^2r_n, \qquad (1-47)$$

其中

$$|r_n| \leqslant |f_{yy}(x_n, y(x_n))| \cdot C_1^2 + \frac{1}{6}|y'''(\xi)| \leqslant C_2,$$

C_2 为另一个常数．定义函数

$$g(x) = f_y(x, y(x)).$$

我们将（1-47）看成把 Euler 方法应用于函数 $e(x)$ 的一个新微分方程

$$e'(x) = g(x)e(x) - \frac{1}{2}y''(x) \qquad (1-48)$$

求解的结果，而在每一步上取不超过 h^2C_2 的一个附加误差．因为 $e_0 = 0$，初值 $\bar{e}_0$ 为零．对这个方程我们可应用定理 1.4，便有以下结论．

定理 1.5． 令函数 $f(x, y)$ 在区域 R 内（见 1.2-4）有连续的一阶和二阶导数，则对 $y' = f(x, y)$，$y(a) = \eta$ 的解 $y(x)$ 的 Euler 近似误差 e_n 可写成形式

$$e_n = he(x_n) + O(h^2), \qquad (1-49)$$

其中 $e(x)$ 是

$$e'(x) = f_y(x, y(x))e(x) - \frac{1}{2}y''(x) \qquad (1-50)$$

的解，且满足 $e(a) = 0$．更为确切地说，

$$\left|\frac{1}{h}e_n - e(x_n)\right| \leqslant hC_3 \cdot E_{L_1}(x_n - a), \qquad (1-51)$$

其中

$$C_3 = \frac{1}{2}\max_{a \leqslant x \leqslant b}|e''(x)| + C_2,$$

$$L_1 = \max_{a \leqslant x \leqslant b}|f_y(x, y(x))|.$$

常数 C_3 的界可用 $f(x, y)$ 的导数的界来表示（见本章末问题 7）．由（1-50）确定的函数 $e(x)$ 称为这个问题的伸缩误差函

数.

1.3-4. 应用. 我们对前面考察的二个例子来应用定理1.5的结果. 在第一个例子中，$y'=y$，有 $y(x)=e^x$，从而

$$f_y(x,y(x))=1,\quad y''=e^x.$$

$e(x)$的方程是

$$e'(x)=e(x)-\frac{1}{2}e^x,\quad e(0)=0,$$

其解为

$$e(x)=-\frac{1}{2}xe^x.$$

于是定理 1.5 给出

$$e_n=-\frac{1}{2}x_ne^{x_n}h+O(h^2).$$

在第二个例中，$y'=-y$，其真解为 $y(x)=e^{-x}$. 因此

$$f_y(x,y(x))=-1,\quad y''=e^{-x}.$$

伸缩误差函数满足

$$e'(x)=-e(x)-\frac{1}{2}e^{-x},\quad e(0)=0,$$

从而给出

$$e(x)=-\frac{1}{2}xe^{-x}. \tag{1-52}$$

于是利用定理 1.5，有

$$e_n=-\frac{1}{2}x_ne^{-x_n}h+O(h^2). \tag{1-53}$$

特别是，对于 $x_n=1$，$n=1/h$，

$$\lim_{h\to 0}h^{-1}e_n=-\frac{1}{2}e^{-1},$$

这与表 1.3 的值是一致的.

在表 1.4 中，我们对于 $y'=-y$ 把对应于 $h=1/64$ 的真

实误差 e_n 与 (1-52) 所给出的值 $he(x_n)$ 相比较，可以看到函数 $e(x)$ 很精确地反映出真实误差的性态.

当然，在实际问题中，真解是不知道的，并且函数 $e(x)$ 也不易确定. 但是，以后我们将看到，尽管那样，还存在确定 $e(x)$ 的实用方法并且能更好地使用类似于 (1-49) 的公式.

表 1.4　真实误差与用伸缩误差函数计算的误差相比较

x_n	1	2	3	4	5
e_n	−0.002892	−0.002120	−0.001165	−0.000570	−0.000261
$he(x_n)$	−0.002874	−0.002114	−0.001167	−0.000572	−0.000263

即使在 $e(x)$ 是未知的情形，只要 (1-49) 成立也是有用的. 我们不使用通常的记号，而用 $y(x, h)$ 表示取步长为 h (必须能整除 $x-a$) 在点 x 上所得到的近似解，因而

$$y(x,h)=y_{(x-a)/h}.$$

从而关系式 (1-49) 可写成

$$y(x,h)=y(x)+he(x)+O(h^2). \qquad (1\text{-}54)$$

如果对两个不同的 h 值，比如说 h 和 qh，计算出 $y(x,h)$，其中 q 与 $\frac{1}{2}$ 成比例，那么我们从二个关系式 (1-54) 及

$$y(x,qh)=y(x)+qhe(x)+O((qh)^2),$$

可消去未知值 $e(x)$. 由于 $O(h^2)+O((qh)^2)=O(h^2)$，结果是

$$y(x)=\frac{y(x,qh)-qy(x,h)}{1-q}+O(h^2). \qquad (1\text{-}55)$$

在 (1-55) 中的项 $O(h^2)$ 当然仅在极少的情形才为零(甚至对我们所考察的例 $y'=\pm y$ 也不为零). 然而，如果 h 甚小，由 (1-55) 中略去 $O(h^2)$ 而得到的值却明显地要比 $y(x,h)$ 或 $y(x,hq)$ 为好. 我们用表 1.5 来说明这一点，表 1.5 是对表 1.3

给出的值 y_n 应用(1-55)而得到的，最后两个的值总是可以使用的 $\left[q=\frac{1}{2}\right]$.

上述外推方法为 L. F. Richardson 首次广泛地使用，并且他称之为"延迟趋向于极限"。形象地说，我们称它外推至 $h=0$ 或简称 h-外推。这个思想出现在数值分析许多其它分支中，当依赖于一个参数的数值方法中的误差有简单的渐近性态时，便可应用这个方法。

表 1.5 延迟趋向于极限

k	$y(1,2^{-k})$	改进值
4	0.356074	
5	0.362055	0.368036
6	0.364987	0.367919
7	0.366438	0.367889
8	0.367160	0.367882
⋮		
∞	0.367879	

1.4. Euler 方法的舍入误差

1.4-1. *局部舍入误差*. 除了参与计算的所有参数都是有理数并且全都使用有理运算这种不大可能的情形外，定义 Euler 方法的关系式(1-1)绝不能精确地被满足，因为有舍入误差。研究舍入误差并不简单，因为通常的运算规则不再成立。为了使这种讨论建立在牢固的基础上，就有必要很仔细和明确地陈述基本假设。

除另作申明外，我们将假设所有算术运算都是按定点运算来完成的，从而在计算中进行的所有数都是机器可处理的最小正数的整数倍数。用 u 表示这个数，并称为机器的基本

单位[1]. 如果 x 是属于机器范围内的任意一个数，那么用 x^* 表示正确舍入后的机器所表示的 x 值. "正确舍入"这句话是指

$$|x - x^*| \leqslant \frac{1}{2} u.$$

总是假设步长 h 和初值点 a 都是准确的机器数，即是 $h^* = h$, $a^* = a$. 这就是说，节点 x_n 都可准确地计算. 为了简化，我们还假设 $y_0^* = y_0$, 虽然这个假设不是实质的.

在这些假设下，用以下关系式来表示真实计算出的值 $\tilde{y}_n$:

$$\tilde{y}_0 = y_0,$$

$$\tilde{y}_{n+1} = \tilde{y}_n + (h\tilde{f}(x_n, \tilde{y}_n))^*, \quad n = 0, 1, 2, \cdots, \qquad (1\text{-}56)$$

这里 $\tilde{f}(x_n, \tilde{y}_n)$ 表示 $f(x_n, \tilde{y}_n)$ 的一个近似值. 理想地，$\tilde{f} = f^*$; 但是我们必须允许这种可能性，即不能十分精确计算出函数 $f(x, y)$, 例如因为在计算中常常有舍入误差的累积，或者因为使用不精确的子程序.

(1-56) 的第二个方程不便于做分析工作，因此我们用以下关系来代替它：

$$\tilde{y}_{n+1} = \tilde{y}_n + hf(x_n, \tilde{y}_n) + \varepsilon_{n+1}. \qquad (1\text{-}57)$$

由这个方程所确定的量 ε_{n+1} 称为局部舍入误差. 于是

$$\varepsilon_{n+1} = (h\tilde{f}(x_n, \tilde{y}_n))^* - hf(x_n, \tilde{y}_n).$$

为了估计 ε_{n+1} 的大小，令

$$\varepsilon_{n+1} = \pi_{n+1} + \rho_{n+1},$$

其中

$$\pi_{n+1} = (h\tilde{f}(x_n, \tilde{y}_n))^* - h\tilde{f}(x_n, \tilde{y}_n),$$

$$\rho_{n+1} = h[\tilde{f}(x_n, \tilde{y}_n) - f(x_n, \tilde{y}_n)],$$

量 π_{n+1} 称为引入误差，它是由乘积 $h\tilde{f}$ 舍入引起的，并且可以

---

1) 对于 IBM704, 709 和 7090 机，如果采用定点运算，有 $u = 2^{-35}$.

像 $\frac{1}{2}u$ 一样大．量 ρ_{n+1} 称为固有误差，它是由对函数 f 计算的不精确性所引起的．即使这个不精确性为最低限度的有效数字中的几个单位，而固有误差的阶却为 hu，于是在所考虑的情形下，它比引入误差小得多．

1.4-2．累积舍入误差的一个先验界． 在 1.4 剩下的部分中，在

$$|\varepsilon_n| \leqslant \varepsilon, \quad n = 1, 2, \cdots \tag{1-58}$$

的单独假设下，其中 ε 是一个常数，我们将涉及到估计累积舍入误差 $r_n = \tilde{y}_n - y_n$ 的问题．如果只出现引入误差，那么可取 ε 为 $\frac{1}{2}u$．在存在固有误差时，这个界还是相当精确的，因为 f 的任何误差的影响用小的数 h 相乘后就大大地减弱．现在我们使用证明定理 1.3 的类似方式来进行分析． 从(1-57)减去差分方程的精确解 y_n 满足的对应关系式，并令

$$r_n = \tilde{y}_n - y_n,$$

得到

$$r_0 = 0,$$

$$r_{n+1} = r_n + h[f(x_n, \tilde{y}_n) - f(x_n, y_n)] + \varepsilon_{n+1},$$

$$n = 0, 1, 2, \cdots.$$

利用 Lipschitz 条件以及 (1-58)，从而导出

$$|r_{n+1}| \leqslant (1 + hL)|r_n| + \varepsilon, \ n = 0, 1, \cdots,$$

用引理 1.2 可以求解这个递推关系式． 我们把这个结果陈述为以下定理：

定理 1.6． 如果 $f(x,y)$ 满足条件(B)，并且局部舍入误差满足(1-58)，那么从准确值开始的 Euler 方法的累积舍入误差 r_n 服从

$$|r_n| \leqslant \frac{\varepsilon}{h} E_L(x_n - a), \quad n = 0, 1, \cdots. \tag{1-59}$$

如果 ϵ 是半个基本单位，(1-59) 表明如下事实，在 n 步后累积舍入误差至多是基本单位的 $(2h)^{-1}E_L(x_n-a)$ 倍． 对于 $L=0$，这个表达式化成 $(2h)^{-1}(x_n-a)=\frac{1}{2}n$． 这个结果容许有一个简单解释： $L=0$ 意即 $f(x,y)$ 与 y 无关． 用 Euler 方法于方程

$$y'=f(x),\quad y(a)=\eta$$

的解，便形成和 $\sum_{m=0}^{n-1}hf(x_m)$，并且舍入误差显然是以基本单位的 $\frac{n}{2}$ 倍为界．

对于初值问题 $y'=y$，$y(0)=1$，有 $L=1$，并且上面的结果给出

$$|r_n|\leqslant\frac{u}{2h}(e^{x_n}-1).$$

特别是，对于 $h=10^{-3}$，$n=1000$，我们求得

$$|r_{1000}|\leqslant\frac{1}{2}u\cdot10^3(e-1)=859u.$$

对于 $y'=-y$ 仍得相同的结果，因为在这种情形下，我们也有 $L=1$．

1.4-3．*累积舍入误差对局部舍入误差的依赖性．* 象离散误差情形一样，假设准确解 $y(x)$（至少近似地）是已知的，便可得出更加真实的估计．作为一种准备，我们将导出可显示出 r_n 依赖于局部舍入误差的一个公式． 为了不立即对付太多的困难，在本章我们仅考察线性方程[1]

$$y'=g(x)y+p(x),\tag{1-60}$$

这里假设 $g(x)$ 与 $p(x)$ 在区间 $[a,b]$ 内是连续的．对于这种

1) 在第二章中讨论一般微分方程．

情形，理论上的近似为

$$y_{n+1}=y_n+h[g(x_n)y_n+p(x_n)]; \tag{1-61}$$

数值近似 $\tilde{y}_n$ 满足

$$\tilde{y}_{n+1}=\tilde{y}_n+h[g(x_n)\tilde{y}_n+p(x_n)]+\varepsilon_{n+1}. \tag{1-62}$$

从 (1-62) 减去 (1-61)，求得累积舍入误差 r_n 的关系式

$$r_{n+1}=r_n+hg(x_n)r_n+\varepsilon_{n+1}. \tag{1-63}$$

我们把最后的关系式看成是变量 r_n 的差分方程，并在初始条件 $r_0=0$ 下求解．我们寻求这个方程具有如下形式的解：

$$r_n=\sum_{m=1}^{n} d_{n,m}\varepsilon_m, \tag{1-64}$$

它显示出 r_n 对所有前面的局部舍入误差 ε_m 的依赖性．为确定常数 $d_{n,m}$，代入 (1-63)：

$$\sum_{m=1}^{n+1} d_{n+1,m}\varepsilon_m=\sum_{m=1}^{n} d_{n,m}\varepsilon_m+hg(x_n)\sum_{m=1}^{n} d_{n,m}\varepsilon_m+\varepsilon_{n+1},$$

不论单个舍入误差 ε_m 如何，都必须满足这个关系式．比较两边 $\varepsilon_m(m=1,2,\cdots,n)$ 的系数，我们得到

$$d_{n+1,m}=d_{n,m}+hg(x_n)d_{n,m}$$
$$(n=0,1,\cdots;m=1,2,\cdots,n). \tag{1-65a}$$

比较 ε_{n+1} 的系数，结果是

$$d_{n+1,n+1}=1 \quad (n=0,1,\cdots). \tag{1-65b}$$

反之，如果系数 $d_{n,m}$ 为 (1-65) 所确定，并且 r_n 为 (1-64) 所确定，那么容易看出，r_n 满足 (1-63).

让我们对方程 $y'=y$ 应用这个结果．在这种情形，

$$g(x)=1,$$

并求得 (1-65) 的解为

$$d_{n,m}=(1+h)^{n-m} \quad (n\geq m),$$

从而

$$r_n = \sum_{m=1}^{n} (1 + h)^{n-m} \varepsilon_m.$$

一般地，方程组（1-65）不能显式求解．在损失一些精确度的条件下，我们能得到 r_n 的一个显示表达式．在（1-65a）中的下标 m 保持不变，它只依赖于 n，我们认出这个方程可看成是对未知函数 $d_m(x)$ 的微分方程

$$d'_m(x) = g(x) d_m(x) \tag{1-66a}$$

应用 Euler 方法的结果．关系式（1-65b）提供必要的初值条件

$$d_m(x_m) = 1, \tag{1-66b}$$

这个初值问题可以显式地求解．令

$$G(x) = \int_a^x g(t) dt,$$

则其解为

$$d_m(x) = e^{G(x) - G(x_m)}. \tag{1-67}$$

利用定理 1.3，量 $d_{n,m}$ 与 $d_m(x_n)$ 之差小于 Ch，其中 C 是不依赖于 h, n 或 m 的一个常数．于是

$$|d_{n,m} - \exp[G(x_n) - G(x_m)]| \leqslant Ch,$$
$$a \leqslant x_m \leqslant x_n \leqslant b.$$

把（1-64）中的 $d_{n,m}$ 换为 $d_m(x_n)$，并且估计所涉及到的误差，我们求得

$$r_n = \sum_{m=1}^{n} \exp[G(x_n) - G(x_m)] \varepsilon_m + \theta h C \sum_{m=1}^{n} |\varepsilon_m|, \tag{1-68}$$

其中 $-1 \leqslant \theta \leqslant 1$．如果我们不考虑修正项，那么这个方程表明了局部误差 ε_m 对在后面点上的累积舍入误差 r_n 的影响是增大的，当且仅当 $G(x_n) - G(x_m) > 0$，即是，如果

$$\int_{x_m}^{x_n} g(t)dt > 0.$$

1.4-4. 一个改进的界. 在假设

$$|\varepsilon_m| \leqslant \varepsilon \quad (m = 1,2,\cdots) \tag{1-69}$$

下，为得到对 $|r_n|$ 的一个改进的界，方程 (1-64) 是有用的. 把 (1-65a) 写成形式

$$d_{n+1,m} = (1 + hg(x_n))d_{n,m},$$

我们立即看出，由于 $d_{m,m} = 1$，所有系数 $d_{n,m}$ 都为正. 如果 $h < h_0$，其中

$$h_0 = \begin{cases} \infty, & \text{如果 } g(x) \geqslant 0, x \in [a,b]; \\ \min\limits_{x \in [a,b]} (-g(x))^{-1}, & \text{其它.} \end{cases}$$

于是，如果 (1-69) 成立并且 $h < h_0$，

$$|r_n| \leqslant \varepsilon \sum_{m=1}^{n} d_{n,m}.$$

为了用封闭形式来表达最后的和，令

$$m_n = h \sum_{m=1}^{n} d_{n,m}. \tag{1-70}$$

利用 (1-65)，我们求得

$$\begin{aligned} m_{n+1} - m_n &= h\left(\sum_{m=1}^{n+1} d_{n+1,m} - \sum_{m=1}^{n} d_{n,m}\right) \\ &= h\left(d_{n+1,n+1} + \sum_{m=1}^{n} (d_{n+1,m} - d_{n,m})\right) \\ &= h\left(1 + \sum_{m=1}^{n} hg(x_n)d_{n,m}\right) \\ &= h[1 + g(x_n)m_n]. \end{aligned}$$

于是

$$m_{n+1} = m_n + h[g(x_n)m_n + 1].$$

显然，$m_0=0$. 利用定理 1.3，故

$$m_n=m(x_n)+\theta_n Ch \quad (|\theta_n|\leqslant 1), \tag{1-71}$$

其中函数 $m(x)$ 是初值问题

$$m'(x)=g(x)m(x)+1, \tag{1-72}$$
$$m(a)=0$$

的解，并且 C 是仅依赖于所给定的微分方程的一个常数. 于是我们证明了：

定理 1.7. 如果用 Euler 方法得到 (1-60) 解的局部舍入误差满足 (1-69)，并且 $h<h_0$，那么累积误差满足

$$|r_n|\leqslant\frac{\varepsilon}{h}\{m(x_n)+O(h)\}, \tag{1-73}$$

其中 $m(x)$ 由 (1-72) 确定.

对于问题 $y'=y$, $y(0)=1$，我们求得 $m(x)=e^x-1$，假设 $\varepsilon=\frac{u}{2}$，那么

$$|r_n|\leqslant\frac{u}{2h}[(e^{x_n}-1)+O(h)].$$

如果略去 $O(h)$ 项，对于 $h=10^{-3}$, $x_n=1$，从而导出

$$|r_{1000}|\leqslant\frac{1}{2}u\cdot 10^3(e-1)=859u,$$

这并未改进上面定理 1.6 的先验界.

另一方面，如果求解的初值问题是 $y'=-y$, $y(0)=1$，我们有 $m(x)=1-e^{-x}$，并且取与上面相同的数据，那么

$$|r_{1000}|\leqslant\frac{1}{2}u\cdot 10^3(1-e^{-1})=326u,$$

这对上面已有结果是一个改进. 考虑大的 x_n 值时，这个差别就更加显著.

1.5. 随机变量

上面给出的累积舍入误差界是在这样保守的假设下导出的，即所有舍入误差为固定符号且取最大值，从而有系统地相互增强．这样一个有规律的误差分布法则，虽然理论上是可能的，而在实际上是绝不可能发生的．界（1-73）尽管理论上“最好可能”，但它却大大地超过真实的误差．因而希望从另一方面对舍入误差进行估计，就是关于舍入误差用“平均”或“正常”增长来描述．把局部舍入误差看成随机变量便可得到这样的估计．

并不认为读者具有随机变量的任何知识．在讨论把它们应用到舍入误差之前，我们先介绍与它们有关的一些基本知识[1]．

1.5-1．随机变量的定义；分布． 在未作进一步介绍前，用记号 ξ 表示随机变量，它是在一致的条件下进行重复多次的随机试验的数值结果．掷骰子便是古典的一个例子．我们假设 ξ 与概率分布有关．根据我们的目的，如下定义这个概念就够了． 令任意一个区间 I 是给定的，并且进行 N 次试验，其中 N 为大数；令 N_I 是试验结果为 ξ 的次数，而 ξ 是属于 I 的一个数．如果

$$\lim_{N\to\infty}\frac{N_I}{N}=P(\xi\in I) \tag{1-74}$$

存在，那么称它为 ξ 属于 I 的概率．如果对整个区间 I 这个极限存在，那么便说变量 ξ 具有概率分布．

---

1) 这里提出的理论只是些概要． 对它更为详细的论述，读者可参阅标准教程，例如(从易到难) Hoel [1954]，Feller [1957]，Cramér [1946] 及 Loéve [1955].

对于形如 $a < x \leqslant b$ 的区间 I，概率 $P(\xi \in I)$ 可用单个变量函数来表示．如果我们规定

$$F(x) = P(\xi \leqslant x), \tag{1-75}$$

那么显然有

$$\begin{aligned} P(a < \xi \leqslant b) &= P(\xi \leqslant b) - P(\xi \leqslant a) \\ &= F(b) - F(a). \end{aligned}$$

函数 $F(x)$ 称为变量 ξ 的分布函数．根据概率的定义，$F(x)$ 是一个非减函数，它的定义域为实轴，值域为区间 $[0,1]$．

如果分布函数 $F(x)$ 是连续且逐段连续可微的，那么称导数 $p(x) = F'(x)$ 为随机变量的频率函数或概率密度．的确，由于

$$\lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = p(x),$$

所以 ξ 位于 x 和 $x + \Delta x$ 之间的概率近似地为 $p(x)\Delta x$．而且，如果 $\lim\limits_{x \to -\infty} F(x) = 0$，那么

$$\begin{aligned} P(\xi \leqslant x) = F(x) &= \int_{-\infty}^{x} \frac{dF(t)}{dx} dt \\ &= \int_{-\infty}^{x} P(t)\, dt. \end{aligned} \tag{1-76a}$$

在许多应用中（例如掷骰子），随机变量 ξ 仅假设为（有限个或无限多个）离散值 $x_\nu (\nu = 0, 1, 2, \cdots)$[1]．令 P_ν 是 ξ 取值为 x_ν 的概率，那么分布函数在任何两个相邻点 x_ν 之间为常数，并且在点 x_n 有跳跃 $\Delta F_\nu = P_\nu$．从而 $\xi \leqslant x$ 的概率为

$$P(\xi \leqslant x) = \sum_{x_\nu \leqslant x} \Delta F_\nu = \sum_{x_\nu \leqslant x} P_\nu. \tag{1-76b}$$

如果采用 Stieltjes 积分记号，那么 (1-76a) 和 (1-76b) 都可统一为一个公式

1) 这些值与上面讨论的节点无关．

$$P(\xi \leqslant x) = \int_{-\infty}^{x} dF(x). \tag{1-76}$$

由于在本书中仅考虑刚刚所讨论的二种类型的分布函数，对于我们的问题不论采用由(1-76a)或(1-76b)来规定(1-76)都是可以的。

例．(a) 假设 ξ 仅取区间 $[a,b]$ 内的值，并且假设在这个区间内所有的值都有相等的概率，那么分布函数为

$$F(x) = \begin{cases} 0, & x \leqslant a, \\ \dfrac{x-a}{b-a}, & a \leqslant x \leqslant b, \\ 1, & b \leqslant x. \end{cases}$$

概率密度在 $[a,b]$ 外为零；在 $[a,b]$ 内为定值

$$p = \frac{1}{b-a}.$$

这样的分布称为矩形分布(见图 1.3)。

(b) 离散分布的一个简单的例子便是掷骰子出现的点

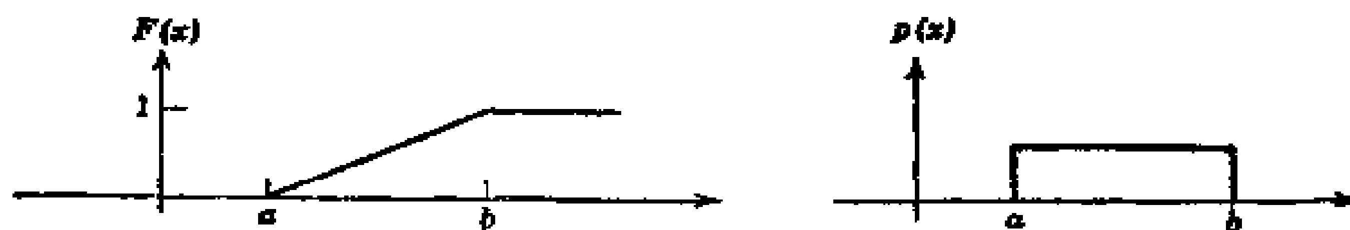


图 1.3



图 1.4

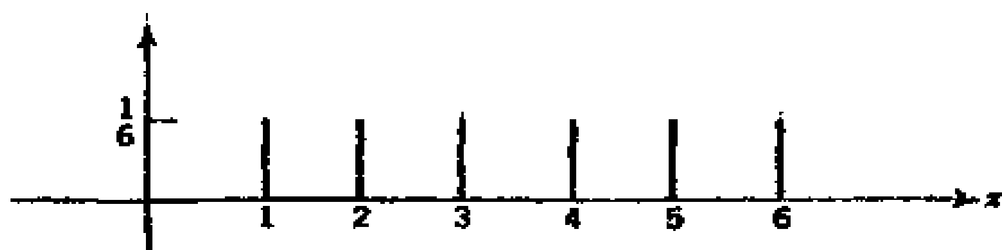


图 1.5

数．这里 ξ 仅取6个值 $x_\nu=\nu$，$\nu=1,2,\cdots,6$ 中的一个．对于一个理想骰子，$P_\nu=\frac{1}{6}$．分布函数如图1.4所表示的图形．在这种情形，概率密度不存在．但是，概率 p_ν 可以用在点 x_ν 的垂直线段长为 p_ν 的图形来表示(见图1.5)．

1.5-2．**随机变量的均值与方差．** 可能有多种类型的分布函数；在右端连续并且满足 $F(-\infty)=0$ 和 $F(\infty)=1$ 的每一个非减函数 $F(x)$ 都可定义为某一个随机变量的分布函数．然而，在许多实际问题中，只用两个参数便足以表征出分布函数，这两个参数称为分布的均值与方差．

随机变量 ξ 的均值 μ 是进行 N 次试验的平均结果当 $N\to\infty$ 时的极限．于是，如果 ξ_i 是第 i 次试验的结果，那么

$$\mu=\lim_{N\to\infty}\frac{1}{N}(\xi_1+\xi_2+\cdots+\xi_N).\qquad(1\text{-}77)$$

我们希望把 μ 用分布函数来表示，并且在离散概率分布的情形，作如下的讨论．如果 ξ_i 取值 x_ν，便把它放在 B_ν 盒内，对于大的 N 值，根据概率的定义，每一个 B_ν 盒内包含有数 ξ_1，$\xi_2,\cdots\xi_N$ 的个数为 $N(p_\nu+\delta_\nu)$，其中 $\delta_\nu\to0$，当 $N\to\infty$ 时．由于每一个 ξ_i 只在一个盒内，对所有盒内的基值求和，我们有

$$\frac{1}{N}(\xi_1+\xi_2+\cdots+\xi_N)=\sum_\nu x_\nu(p_\nu+\delta_\nu).$$

于是，令 $N\to\infty$，

$$\mu=\sum_\nu x_\nu p_\nu.\qquad(1\text{-}78\text{a})$$

在连续分布的情形，采用类似但更为精确的论证，我们求得

$$\mu = \int_{-\infty}^{\infty} x p(x) dx. \quad (1\text{-}78b)$$

(1-78a) 与 (1-78b) 可统一成方程

$$\mu = \int_{-\infty}^{\infty} x dF(x). \quad (1\text{-}78)$$

给出上面的启发性的解释之后，可将它作为均值 μ 的定义。通常我们令

$$\mu = E(\xi),$$

并且称 $E(\xi)$ 为随机变量 ξ 的期望值。对于上面的例子，我们容易求得，对 (a)，$E(\xi) = \frac{1}{2}(a+b)$；对 (b)，$E(\xi) = \frac{7}{2}$。

如果 ξ 是任意的随机变量，并且 $f(x)$ 是 x 的任意函数，那么 $f(\xi)$ 仍是随机变量。如果它的期望值是存在的，我们利用在 $\xi = x$ 的概率乘 $f(x)$ 并且对所有 x 值求和，便得到这个期望值。于是，在离散情形，

$$E(f(\xi)) = \sum_{\nu} f(x_{\nu}) p_{\nu}; \quad (1\text{-}79a)$$

在连续情形，

$$E(f(\xi)) = \int_{-\infty}^{\infty} f(x) p(x) dx. \quad (1\text{-}79b)$$

这两种情形都可统一成方程

$$E(f(\xi)) = \int_{-\infty}^{\infty} f(x) dF(x). \quad (1\text{-}79)$$

作为这些法则的一个应用，我们注意到，如果 a 和 b 都是任意常数，那么

$$E(a\xi) = \int_{-\infty}^{\infty} ax dF(x) = aE(\xi) \quad (1\text{-}80)$$

及

$$E(\xi + b) = \int_{-\infty}^{\infty} (x+b) dF(x) = E(\xi) + b. \quad (1\text{-}81)$$

方程 (1-78a) 容许有简单力学的解释. 如果把质量为 p_ν 的质点放在 x 轴的点 $x = x_\nu$ 上，那么，由于 $\Sigma_\nu p_\nu = 1$，所有质点的公共重心显然位于点 $x = \mu$ 上. 在连续分布情形，类似解释也是可能的. 因此期望值可看成对随机变量位置的粗糙的度量. 但是，并未说明关于 ξ 的展形或离差. 恒为零的随机变量均值为零，以及假设取 $\pm 10^6$ 的每一个值具有概率为 $\frac{1}{2}$ 的随机变量，其均值为零. 在很多场合，为了方便，引入称为 ξ 的方差的一个非负量作为对离差的度量，并且规定它为

$$\operatorname{var}(\xi) = E((\xi - \mu)^2). \tag{1-82}$$

因此方差是 ξ 与它的均值 μ 的偏差平方的期望值. 利用上面的公式，取 $f(\xi) = (\xi - \mu)^2$，我们求得显示表达式，在离散情形，

$$\operatorname{var}(\xi) = \sum_\nu (x_\nu - \mu)^2 p_\nu; \tag{1-83a}$$

在连续情形，

$$\operatorname{var}(\xi) = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx. \tag{1-83b}$$

这二个公式都是

$$\operatorname{var}(\xi) = \int_{-\infty}^{\infty} (x - \mu)^2 dF(x) \tag{1-83}$$

的特殊情形. 对于上面所考虑的特殊分布，我们求得，对 (a)，

$$\operatorname{var}(\xi) = \frac{1}{b-a}\int_a^b \left(x - \frac{b+a}{2}\right)^2 dx = \frac{(b-a)^2}{12};$$

对 (b)，

$$\operatorname{var}(\xi) = \frac{1}{6}\left[\left(-\frac{5}{2}\right)^2 + \left(-\frac{3}{2}\right)^2 + \left(-\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{3}{2}\right)^2 + \left(\frac{5}{2}\right)^2\right] = \frac{35}{12}.$$

我们还要注意到，从方差的定义和从 (1-80) 及 (1-81)

可直接导出以下的关系：如果 a 及 b 是任意二个常量，那么

$$\operatorname{var}(a\xi)=a^2\operatorname{var}(\xi), \tag{1-84}$$

$$\operatorname{var}(\xi+b)=\operatorname{var}(\xi). \tag{1-85}$$

在力学上方差相当于围绕上面所考虑的质点系重心的惯量矩。

如果 $\operatorname{var}(\xi)=\sigma^2$，那么非负量 σ 与 ξ 和 μ 有相同的量纲。称它为随机变量 ξ 的标准偏差，并且此刻可作为对随机变量与其期望值的"平均"偏差的度量。σ 的深刻意义与随机变量的和连系起来就变得明显了。

1.5-3．多个随机变量的函数．令 ξ 与 η 是由同时进行两个试验中所引起的二个随机变量，并令它们的分布函数分别是 $F(x)$和 $G(y)$．我们还引入同时使 $\xi\leqslant x$ 和 $\eta\leqslant y$ 成立的作为概率的 ξ 及 η 的联合分布函数 $F(x,y)$：

$$F(x,y)=P(\xi\leqslant x \,\&\, \eta\leqslant y).$$

如果 $f(x,y)$ 是 x 与 y 的任意数值函数，那么 $f(\xi,\eta)$ 仍是一个随机变量．我们用 $\xi=x,\eta=y$ 的概率乘 $f(x,y)$，并对所有 x 和 y 求和，使得到它的期望值，这个结果可写成一般形式

$$E(f(\xi,\eta))=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)dF(x,y). \tag{1-86}$$

对特殊情形 $f(\xi,\eta)=\xi+\eta$ 及 $f(\xi,\eta)=\xi\eta$，我们来计算这个积分．

我们注意到，对于 $f(\xi,\eta)=\xi$，积分(1-86)可给出变量 ξ 的期望值，另一个变量 η 则完全不予考虑，从而

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x\,dF(x,y)=E(\xi).$$

类似地，

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}y\,dF(x,y)=E(\eta).$$

由此立即导出基本结果

$$E(\xi+\eta)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)dF(x,y)$$
$$=E(\xi)+E(\eta). \tag{1-87}$$

于是它可陈述为：和的期望值等于期望值的和.

对于两个随机变量的乘积，一般说来并没有这样简单法则成立. 幸而，如果两个随机变量是独立的，则有简单的结果成立. 如果对于任何二个区间 I 和 J，同时使得 $\xi\in I$ 和 $\eta\in J$ 的概率等于 $\xi\in I$ 和 $\eta\in J$ 的单个概率的乘积，则称变量 ξ 与 η 是独立的. 在实际应用中，如果规定 ξ 与 η 的两个试验是互不干扰的，则这二个随机变量便可看成是独立的. 使用概率的语言，独立的条件可写成

$$P(\xi\in I\ \&\ \eta\in J)=P(\xi\in I)P(\eta\in J).$$

如果把它应用于区间 $I=[-\infty,x]$，$J=[-\infty,y]$，对于两个独立的随机变量联合分布函数，我们求得基本关系式

$$F(x,y)=F(x)G(y).$$

用最后恒等式，便能对 $f(\xi,\eta)=\xi\eta$ 计算(1-86). 我们求得

$$E(\xi\eta)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xydF(x,y)=\int_{-\infty}^{\infty}xdF(x)\int_{-\infty}^{\infty}ydG(y)$$
$$=E(\xi)E(\eta). \tag{1-88}$$

换言之：二个独立的随机变量的期望值等于期望值的乘积.

我们能够计算二个独立随机变量的和的方差. 根据定义

$$\operatorname{var}(\xi+\eta)=E((\xi+\eta-E(\xi+\eta))^2).$$

引入新的随机变量

$$\xi'=\xi-E(\xi),\quad \eta'=\eta-E(\eta),$$

我们有

$$E(\xi')=0,\quad E(\xi'^2)=\operatorname{var}(\xi),$$
$$E(\eta')=0,\quad E(\eta'^2)=\operatorname{var}(\eta).$$

利用 (1-87) 及 (1-88)，从而

$$\mathrm{var}(\xi+\eta)=E((\xi'+\eta')^2)=E(\xi'^2+2\xi'\eta'+\eta'^2)$$
$$=E(\xi'^2)+2E(\xi')E(\eta')+E(\eta'^2),$$

最后

$$\mathrm{var}(\xi+\eta)=\mathrm{var}(\xi)+\mathrm{var}(\eta).\tag{1-89}$$

这就推得二个独立的随机变量的和的方差等于方差的和.

如果变量不是独立的，项 $E(\xi'\eta')$ 一般不为零，称之为二个变量 ξ' 和 η' 的协方差. 利用归纳法，容易看出基本关系式 (1-87)，(1-88) 及 (1-89) 对于任何有限个随机变量都是成立的. 于是，如果 $\xi_1,\xi_2,\cdots,\xi_n$ 是由同时进行试验引起的 n 个随机变量，那么

$$E(\xi_1+\xi_2+\cdots+\xi_n)$$
$$=E(\xi_1)+E(\xi_2)+\cdots+E(\xi_n).\tag{1-90}$$

此外，如果 ξ_i 的任何二个随机变量都是相互独立的，那么

$$E(\xi_1\xi_2\cdots\xi_n)=E(\xi_1)E(\xi_2)\cdots E(\xi_n)\tag{1-91}$$

及

$$\mathrm{var}(\xi_1+\xi_2+\cdots+\xi_n)=\mathrm{var}(\xi_1)+\mathrm{var}(\xi_2)$$
$$+\cdots+\mathrm{var}(\xi_n).\tag{1-92}$$

1.5-4. 大量随机变量的和. 在以后大部分应用中，我们将涉及到形式如下的随机变量:

$$r_n=d_{n1}\xi_1+d_{n2}\xi_2+\cdots+d_{nn}\xi_n,\tag{1-93}$$

其中 $d_{nm}(n=1,2,\cdots;\ m=1,2,\cdots,n)$ 都是常量，ξ_m 都是独立随机变量，其均值与方差为

$$E(\xi_m)=\mu_m,\quad \mathrm{var}(\xi_m)=\sigma_m^2, m=1,2,\cdots.$$

利用前二节的关系式，容易计算出 r_n 的均值与方差. 由于 (1-90) 及 (1-80)，我们求得

$$E(r_n)=d_{n1}\mu_1+d_{n2}\mu_2+\cdots+d_{nn}\mu_n.\tag{1-94}$$

类似地，由 (1-92) 及 (1-84) 得到

$$\mathrm{var}(r_n) = d_{n1}^2\sigma_1^2 + d_{n2}^2\sigma_2^2 + \cdots + d_{nn}^2\sigma_n^2. \qquad (1\text{-}95)$$

值 $E(r_n)$ 和 $\mathrm{Var}(r_n)$ 表达关于随机变量 r_n 的位置和离差的某些信息，但是必须认识到，它们一般地不能完全确定对应于 r_n 的分布函数. 即使 ξ_m 的分布为已知，除去一些简单情形外，显式确定形如（1-93）的随机变量的分布函数是很困难的问题. 在这些情形，如果（1-93）中的项数 n 趋向无穷时而能找到一个十分简单的情形就够满意了. 可以给出这样的情形，在一些弱的条件下，r_n（适当标准化以后）的分布函数趋向一个简单的函数，它完全是与（1-93）中出现的变量 ξ_m 的分布无关. 这就是所谓概率论的中心极限定理的内容. 我们在下面以一个可直接应用于我们的问题[1]的形式来陈述概率论中一个特殊的中心极限定理.

定理 1.8. （极限定理）. 令 $\xi_m(m=1,2,\cdots)$ 是独立的随机变量，其均值为 μ_m，方差为 σ_m^2，并且使得标准变量 $(\xi_m-\mu_m)/\sigma_m$ 一致有界. 令随机变量 r_n 是由（1-93）确定，并且假设

$$\lim_{n\to\infty}\frac{d_{nm}\sigma_m}{[\mathrm{var}(r_n)]^{1/2}} = 0 \quad (m=1,2,\cdots) \qquad (1\text{-}96)$$

对 m 一致成立. 如果 $F_n(x)(n=1,2,\cdots)$ 是标准化随机变量

$$\bar{r}_n = \frac{r_n - E(r_n)}{[\mathrm{var}(r_n)]^{1/2}} \qquad (1\text{-}97)$$

的分布函数，那么

$$\lim_{n\to\infty} F_n(x) = \Phi(x), \qquad (1\text{-}98)$$

其中

$$\Phi(x) = \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt. \qquad (1\text{-}99)$$

---

1) 这个形式取自 Loève [1955, p. 295]. 应当注意，古典形式的极限定理不能直接应用于我们这种情形，因为这里系数 $d_{n,m}$ 依赖于 n 及 m.

这个定理的证明超出本书范围.

函数 $\Phi(x)$称为正态分布函数，它的导数

$$\varphi(x)=\frac{1}{(2\pi)^{1/2}}e^{-\frac{x^2}{2}} \qquad (1-100)$$

称为正态频率函数. 随机变量其分布为正态分布函数的称为正态分布. 实际上，这个极限定理说明了在条件 (1-96) 下，如果变量 r_n 适当地标准化，对充分大的 n 值，它就近似于正态分布. 更明确地说，它能使我们阐明，对于大的 n 和给定的 x 及 y 值，$x<y$，以及对于

$$xs_n \leqslant r_n - m_n \leqslant ys_n,$$

其中 $s_n=(\mathrm{var}(r_n))^{1/2}$，$m_n=E(r_n)$ 的概率近似地为

$$\Phi(y)-\Phi(x).$$

特别是，如果 $x=-y$，$|r_n-m_n|>ys_n$ 的概率近似于

$$\pi(y)=1-(\Phi(y)-\Phi(-y)).$$

在下表中，我们给出 $x=1,2,\cdots,5$ 的函数$\pi(x)$的一些数值.

x	1	2	3	4	5
$\pi(x)$	0.31731	0.04500	0.00270	0.00006	0.00000

由此可见，对于正态分布的变量，它与均值的偏差仅约 32% 超过标准偏差，并且大约只有 27% 超过标准偏差的三倍.

正态分布函数的其它值可以从现存的表中查得（国立标准局 [1942]）.

1.6. 舍入误差的概率理论

回到在 1.4 开始的舍入误差的讨论. 现在我们提出局部

舍入误差是随机变量的假设。

这个假设立即会引起批评。严格地说，假设使用正常运算状态的数字计算设备，用完全确定的方法对给定的初值问题的解是一个完全确定的事件。如果在相同的条件下，进行重复多次的计算，每次结果恰好有相同舍入误差。使用大量劳动，有可能预先断定在每一步上舍入误差是什么。但是，这会引起与概率论中古典试验、掷骰子同样的异议。由初值条件决定的同样为严格确定的事件，至少象牛顿力学定律一样长久被认为是成立的。对运动方程积分，在理论上有可能把掷骰子的结果预先确定为初值条件的函数。如果掷骰子总是不能指出相同的点数，这一定是由于初始条件的改变。我们把掷骰子结果认为是一个随机事件，因为它对初值条件依赖关系是十分复杂的。同样，我们把局部舍入误差当作随机事件，因为它们也以十分复杂的方式依赖于指定给解的初值。在这二种情形，微分方程的解和掷骰子这样的样本空间，即，确定随机变量的所有试验的总数都是由改变初值条件得到的。

1.6-1. *局部舍入误差的分布*. 我们记得在1.4-1中引进的约定与定义，特别是，局部舍入误差可分解为固有和引入误差。如那里所述，对局部舍入误差的影响通常将由引入误差引起，我们暂时完全略去固有误差，并且令 $\varepsilon_m = \pi_m$. 不出现固有误差可由试验来实现，例如取 $f(x,y) = cy$, c 为整数。

至于引入误差 π_m 的分布，由定义，π_m 是由乘积 hf^* 舍入引起的误差。因为 h 和 f^* 都是 u 的整数倍，从而 hf^* 是 u^2 的整数倍。如果 h 和 f^* 都是正的，象通常一样，若一个正数恰好位于二个相邻的机器数的中间就进入，那么乘积（以二进制）的舍入误差便取 $\frac{1}{u} = 2^N$ 个值：

$$-\frac{1}{2}u+u^2,\ -\frac{1}{2}u+2u^2,\cdots,\ \frac{1}{2}u-u^2,\ \frac{1}{2}u \tag{1-101}$$

中的一个值．我们的兴趣是对于形如 hf^* 的一切乘积的舍入误差的分布，其中 h 是一个常数，f^* 是任意的．看来好象能自然地来假设（1-101）的所有误差都有相等概率．不过，一个简单的例子则可指出这是未必成立的．如果 $h=\frac{1}{2}$，在二进制中，hf^* 只可能是

$$0（当 f^* 的最后一位数为 0 时）$$

及

$$\frac{1}{2}u（当 f^* 的最后一位数为 1 时）.$$

如果 f^* 的最后一位的二个值看成概率相同，那么二个舍入误差有相同概率为 $\frac{1}{2}$．于是，在这种情形，（1-101）的 2^N 个值中只出现二个．我们的结论是对固定的 h 和随机的 f^* 的乘积 hf^* 的舍入误差的真实分布依赖于 h.

关于一个固定的数和随机的二进制数的乘积，其舍入误差分布的一般结果如下：

定理 1.9. 如果 $u=2^{-N}$，令 h 为常数，

$$h=h_1 2^{-1}+h_2 2^{-2}+\cdots+h_k 2^{-k},$$

其中 $k\leqslant N$，$h_i=0,1(i=1,\cdots,k-1)$，$h_k=1$．如果 z 是形式如下的随机数：

$$z=z_1 2^{-1}+z_2 2^{-2}+\cdots+z_N 2^{-N},$$

其中 $z_i=0,1(i=1,\cdots,N)$，并且对每一个 $i\geqslant N-k+1$，z_i 取这二个值概率相同，那么乘积 hz 的舍入误差取 2^k 个值：

$$-\frac{1}{2}u+uv,\quad -\frac{1}{2}u+2uv,\cdots,\ \frac{1}{2}u-uv,\ \frac{1}{2}u$$

中的每一个值概率为 2^{-k}，其中 $v = 2^{-k}$.

证. 对于任意 x，象通常一样，用 $[x]$ 表示不超过 x 的最大整数(常常称为 x 的整数部分). 我们用 $\{x\}$ 表示数

$$\{x\} = x - [x],$$

称它为 x 的小数部分. 由定义，

$$0 \leqslant \{x\} < 1$$

并且对于任何整数 m，

$$\{x + m\} = \{x\}. \tag{1-102}$$

乘积 hz 的舍入误差 r 依赖于量

$$q = \{2^N hz\}.$$

如果 $0 \leqslant q \leqslant \frac{1}{2}$，那么，$r = -qu$；如果 $\frac{1}{2} \leqslant q < 1$，那么 $r = (1 - q)u$. 如果对每一个整数 ν，$0 \leqslant \nu < 2^k$，以及对预先指定的值 $z_1, \cdots, z_{N-k}$，恰好存在值 $z_{N-k+1}, \cdots, z_N$ 的一个集合，使得

$$q = \nu v, \tag{1-103}$$

这便证明了这个定理.

通过直接相乘，利用 $h_k = 1$，我们求得

$$\begin{aligned} q = \{ & z_N 2^{-k} + (z_{N-1} + h_{k-1} z_N) 2^{-k+1} \\ & + (z_{N-2} + h_{k-1} z_{N-1} + h_{k-2} z_N) 2^{-k+2} + \cdots \\ & + (z_{N-k+1} + h_{k-1} z_{N-k+2} + \cdots + h_1 z_N) 2^{-1} \}. \end{aligned} \tag{1-104}$$

令

$$\nu v = q_1 2^{-1} + q_2 2^{-2} + \cdots + q_k 2^{-k}, \tag{1-105}$$

其中 q_i 为 0 或 1，并且利用它们为 0 或 1 以及满足关系式

$$z_N = q_k,$$

$$z_{N-1} + h_{k-1} z_N = q_{k-1} + 2m_{k-1},$$

$$z_{N-2} + h_{k-1} z_{N-1} + h_{k-2} z_N = q_{k-2} + 2m_{k-2} - m_{k-1},$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$z_{N-k+1} + h_{k-1}z_{N-k+2} + \cdots + h_1 z_N = q_1 + 2m_1 - m_2$$

的条件来递推地确定 $z_N, z_{N-1}, \cdots, z_{N-k+1}$. 这里

$$m_i (i = 1, \cdots k-1)$$

都是适当选取的整数。把这些 z_i 值代人 (1—104)，利用 (1-105) 及 (1-102)，我们求得

$$\begin{aligned} q = \{ & q_k 2^{-k} + (2m_{k-1} + q_{k-1}) 2^{-k+1} \\ & + (2m_{k-2} + q_{k-2} - m_{k-1}) 2^{-k+2} + \cdots \\ & + (2m_1 + q_1 - m_2) 2^{-1} \} \\ = \{ & q_k 2^{-k} + q_{k-1} 2^{-k+1} + \cdots + q_1 2^{-1} + m_1 \} = \nu v. \end{aligned}$$

这 2^k 个方程 (1-103) 中的每一个方程为值 $(z_{N-k+1}, \cdots, z_N)$ 的 2^k 个集合之一所满足。因为没有一个集合满足二个方程，于是定理证毕。

定理 1.9 指出，x_m 的分布依赖于 h 的最后的不为零的数字。分布函数为

$$F_{u,v}(x) = \begin{cases} 0, & x < \frac{-1}{2} u + uv, \\ \nu v, & x_\nu \leqslant x < x_{\nu+1}, \\ 1, & x \geqslant \frac{1}{2} u, \end{cases} \qquad (1\text{-}106)$$

其中

$$x_\nu = \frac{-1}{2} u + \nu u v \quad (\nu = 1, \cdots, 2^k).$$

注意到函数 $F_{u,v}(x)$ 的图形关于点

$$\left(\frac{1}{2} uv, \frac{1}{2} \right) \text{而不是点} \left(0, \frac{1}{2} \right)$$

为对称，这是由于在所陈述的条件下，进人常常比舍去略多些。对应于 $F_{u,v}(x)$ 的概率图形指出，常数概率 $p_\nu = v$ 放在点 $x = x_\nu$ 上 $(\nu = 1, \cdots, 2^k)$.

剩下来的是确定随机变量 π_m 的期望值及方差. 注意到点 x_ν 及 $x_{v^{-1}-\nu+1}$ 关于点 $x=\frac{1}{2}uu$ 的对称位置，便很容易计算出这二个量. 于是，我们得到

$$\mu = E(\pi_m) = v\sum_{\nu=1}^{v^{-1}} x_\nu = v\sum_{\nu=1}^{\frac{1}{2}v^{-1}} (x_\nu + x_{v^{-1}-\nu+1})$$

$$= \frac{1}{2}uv. \qquad (1\text{-}107)$$

类似地，令 $\lambda = \nu - \frac{1}{2}v^{-1}$，则有

$$\sigma^2 = \operatorname{var}(\pi_m) = v\sum_{\nu=1}^{v^{-1}} \left(x_\nu - \frac{1}{2}uv\right)^2$$

$$= 2v\sum_{\lambda=1}^{\frac{1}{2}v^{-1}} \left(\lambda - \frac{1}{2}\right)^2 u^2v^2.$$

利用公式

$$1^2 + 3^2 + \cdots + (2n-1)^2 = \frac{1}{3}n(2n-1)(2n+1),$$

我们容易求得

$$\sigma^2 = \frac{1}{12}u^2(1-v^2). \qquad (1\text{-}108)$$

对于小的 v 值，函数 $F_{u,v}(x)$的图形逼近函数

$$F_u(x) = \begin{cases} 0, & x \leqslant -\frac{1}{2}u, \\ \frac{1}{2} + x/u, & -\frac{1}{2}u \leqslant x \leqslant \frac{1}{2}u, \\ 1, & x \geqslant \frac{1}{2}u \end{cases} \qquad (1\text{-}109)$$

的图形，它对应于在区间 $\left[-\frac{1}{2}u, \frac{1}{2}u\right]$ 上变量 π_m 的连续、

的匀分布. 容易求得这个分布的期望值与方差为

$$\mu=0, \quad \sigma^2=\frac{1}{12}u^2. \tag{1-110}$$

它们确实是当 $v\to 0$ 时 (1-107) 与 (1-108) 的逼近值. 在实际使用的大部分情形中, 用 $F_u(x)$ 来近似 π_m 的分布已足够.

1.6-2. *累积舍入误差的分布.* 我们对关系式 (1-64) 应用在 1.5-3 和 1.5-4 中提到的随机变量的一般结果,并且假设 $\varepsilon_m=\pi_m$ 以及 π_m 分布为 $F_{u,v}(x)$ 或 $F_u(x)$.

由 (1-93), 我们立得

$$E(r_n)=\sum_{m=1}^{n}d_{n,m}E(\varepsilon_m).$$

因为在许多机器上是按绝对值舍入, 而不是按这个数本身来舍入,因此再假设 $E(\varepsilon_m)=\mu$ 便不真实了. 意即对于负的乘积 hg^*, 我们有 $E(\varepsilon_m)=-\mu$. 于是,我们一般地可以说

$$|E(\varepsilon_m)|\leqslant\mu.$$

如果 h 充分小,并且所有 $d_{n,m}$ 都是正的,那么就有

$$|E(r_n)|\leqslant\mu\sum_{m=1}^{n}d_{nm}=\frac{\mu}{h}m_n, \tag{1-111}$$

其中 m_n 由 (1-70) 确定. 利用 (1-71), 我们求得

$$|E(r_n)|\leqslant\frac{\mu}{h}\{m(x_n)+O(h)\}, \tag{1-112}$$

其中 $m(x)$ 是由 (1-72) 确定的函数. 把这个结果与定理 1.7 相比,我们发现 $|E(r_n)|$ 增长至多为 $\max|r_n|$ 的 $\frac{\mu}{\varepsilon}$ 倍. 对于分布 $F_{u,v}(x)$, 我们有 $\mu/\varepsilon=v$.

如果我们希望用类似的方式来确定 $\operatorname{var}(r_n)$, 那么必须作出随机变量 ε_n 都是独立的重要假设. 要证明这个假设或

把它简化些，看来都是不容易的，并且已经给出例子（Huskey [1949]）帮助证明了，在一些极端的条件下，ε_n 在一定程度上是相关的. 由于这些原因，我们将把我们的假设当作一个有效的假设，这须用(且将用)数值试验来证实.

如果 ε_m 是独立的，$\mathrm{var}(\varepsilon_m)=\sigma^2$，应用（1-94），我们求得

$$\mathrm{var}(r_n)=\frac{\sigma^2}{h}v_n, \tag{1-113}$$

其中

$$v_n=h\sum_{m=1}^{n}d^2_{n,m} \tag{1-114}$$

称为简化方差. 我们试图计算和 v_n，利用对它们所建立的差分方程. 我们有

$$\begin{aligned}v_{n+1}-v_n&=h\left\{\sum_{m=1}^{n+1}d^2_{n+1,m}-\sum_{m=1}^{n}d^2_{n,m}\right\}\\&=h\left\{d^2_{n+1,n+1}+\sum_{m=1}^{n}(d^2_{n+1,m}-d^2_{n,m})\right\}\\&=h\left\{1+\sum_{m=1}^{n}(d_{n+1,m}-d_{n,m})(d_{n+1,m}+d_{n,m})\right\}.\end{aligned}$$

再次使用(1-65)，导出

$$\begin{aligned}v_{n+1}-v_n&=h\left\{1+g(x_n)h\sum_{m=1}^{n}d_{n,m}[2d_{n,m}+hg(x_n)d_{n,m}]\right\}\\&=h\{1+2g(x_n)v_n\}+h^2g(x_n)^2v_n.\end{aligned}$$

从(1-67)，我们知道量 v_n 有界. 值 $g^2(x_n)$ 也有界，因为 $g(x_n)$ 是连续的. 于是可写成

$$v_{n+1}-v_n=h\{1+2g(x_n)v_n\}+O(h^2).$$

对这个差分方程应用定理 1.4，因为 $v_0=0$，从而

$$v_n=v(x_n)+O(h), \tag{1-115}$$

其中函数 $v(x)$ 称为简化方差函数，规定它为

$$v(a)=0,$$
$$v'(x)=2g(x)v(x)+1. \tag{1-116}$$

因此我们求得

$$\operatorname{var}(r_n)=\frac{\sigma^2}{h}\{v(x_n)+O(h)\}. \tag{1-117}$$

从这个结果立即可得到这样定性的结论，r_n 的标准偏差对于"平均"舍入误差来说曾被看成是典型的，其阶为 $h^{-1/2}$，它比由（1-73）给出的最大的理论上可能的舍入误差改进 $h^{1/2}$ 因子。

剩下来的是确定 r_n 的分布。因为变量 π_m 有界，如果我们能证明，当 $n\to\infty$，$nh=x>a$ 时，量

$$D_{n,m}=\frac{d_{n,m}}{(d_{n,1}^2+d_{n,2}^2+\cdots+d_{n,n}^2)^{1/2}}$$

对于 m 一致趋向于零，那么可应用中心极限定理（定理 1.8）的结论。规定

$$G(x)=\int_a^x g(t)dt,$$

（1-116）的显式解写成

$$v(x)=e^{2G(x)}\int_a^x e^{-2G(t)}dt,$$

从而推出 $v(x)>0$。利用（1-67），

$$D_{n,m}=h^{1/2}\frac{d_{n,m}}{v_n^{1/2}}=h^{1/2}\left\{\frac{e^{G(x_n)-G(x_m)}}{[v(x_n)]^{1/2}}+O(h)\right\},$$

显然这就是所需要的结果。

容易看出，为了极限定理的正确性，象（1-96）的假设是必要的。假设值 $d_{n,m}$ 为

$$d_{n,m}=\begin{cases}1, & n=m,\\ 0, & n>m.\end{cases}$$

这相当于局部舍入误差对累积舍入误差的影响无限快地衰减。显然，r_n 的分布总是与 ε_m 分布相同，因此一般来说它不是正态的。

由于推导本节结果并未用到 ε_m 分布的任何特殊性质，因而我们可用通常的方法综合这个结果如下：

定理 1.10. 如果用 Euler 方法解微分方程

$$y' = g(x)y + p(x)$$

的局部舍入误差 ε_m 都是有界的、独立的随机变量有

$$|E(\varepsilon_m)| \leqslant \mu \text{ 和 } \operatorname{var}(\varepsilon_n) = \sigma^2,$$

那么累积舍入误差 r_n 是一个随机变量，且满足

$$\begin{aligned} |E(r_n)| &\leqslant \frac{\mu}{h}\{m(x_n) + O(h)\}, \\ \operatorname{var}(r_n) &= \frac{\sigma^2}{h}\{v(x_n) + O(h)\}, \end{aligned} \tag{1-118}$$

其中 $m(x)$ 及 $v(x)$ 分别由 (1.72) 及 (1-116) 确定。而且标准变量 $[r_n - E(r_n)]/[\operatorname{var}(r_n)]^{1/2}$ 的分布当 $h \to 0$，$nh = x - a$ 时逼近正态分布。

1.6-3. **数值例子.** 我们把上面得到的结果应用于微分方程 $y' = \pm y$. 这些方程都是线性的；而且因为 $g(x) = \pm 1$，所以没有固有误差。于是我们指望 1.6-2 的结果成立。

对于 $y' = y$，我们已经求得

$$m(x) = e^x - 1.$$

积分 (1-116)，容易得到

$$v(x) = \frac{1}{2}(e^{2x} - 1).$$

因此，我们希望近似地有

$$E(r_n) = \frac{\mu}{h}(e^{x_n} - 1),$$

$$\operatorname{var}(r_n) = \frac{\sigma^2}{2h}(e^{2x_n} - 1).$$

特别是，利用连续分布 $F_u(x)$ 所给出的 μ 和 σ^2 值，对于

$$n = 10^3,\ h = 10^{-3},\ x_n = 1,$$

求得

$$E(r_{1000}) = 0,$$

$$\operatorname{var}(r_{1000}) = \frac{u^2 \cdot 10^3}{24}(e^2 - 1) = 266.2u^2.$$

根据中心极限定理，r_{1000} 的分布应当近似于正态分布。

进行以下试验的目的是检验这些结论。用 Euler 方法对 500 个不同的初始条件

$$y_0 = y_{0,q} = \frac{1}{8} + q\Delta$$

求解微分方程 $y' = y$，这里 $q = 0, 1, \cdots, 499$，并且

$$\Delta = \frac{1}{3} \cdot 2^{-13} \quad (\text{带舍入})$$

取 $u = 2^{-28}$。比较数值的终值 $\tilde{y}_{n,q}$ 与从公式

$$(1 + h)^{h^{-1}} = e^{h^{-1}\log(1+h)}$$

$$= e\left\{1 - \frac{1}{2}h + \frac{11}{24}h^2 - \frac{21}{48}h^3 + \cdots\right\}$$

计算出来的理论值

$$y_{n,q} = y_{0,q}(1 + h)^{h^{-1}}.$$

把舍入误差

$$r_{n,q} = \tilde{y}_{n,q} - y_{n,q}$$

记录下来，并且从以下公式计算出试验的期望值与方差

$$E(r_n)_e = \frac{1}{500}\sum_{q=0}^{499} r_{n,q},$$

$$\operatorname{var}(r_n)_e = \frac{1}{500}\sum_{q=0}^{499}(r_{n,q} - E(r_n)_e)^2.$$

对于 $n = 1000$，得到以下数值：

$$E(r_n)_e = -0.2u,\ \operatorname{var}(r_n)_e = 266.7u^2.$$

值 $r_{n,q}$ 的分布如图 1.6 所示. 每一个矩形的高与其底组成的区间值 $r_{1000,q}$ 的个数成正比. 用相同的尺度，还标出正态分布.

预估和试验值之间显然是极为一致的，并且真实分布显然与正态分布[1]类似. 在这种情形，我们断定，舍入误差的统计理论被这个事实证明为正确的.

对于微分方程 $y' = -y$，会得到类似的结果. 对这个方程，求得

$$v(x) = \frac{1}{2}(1 - e^{-2x}),$$

利用和上面相同的数据，于是

$$E(r_{1000}) = 0,$$

$$\text{var}(r_{1000}) = \frac{u^2 \cdot 10^3}{24}(1 - e^{-2}) = 36.0u^2.$$

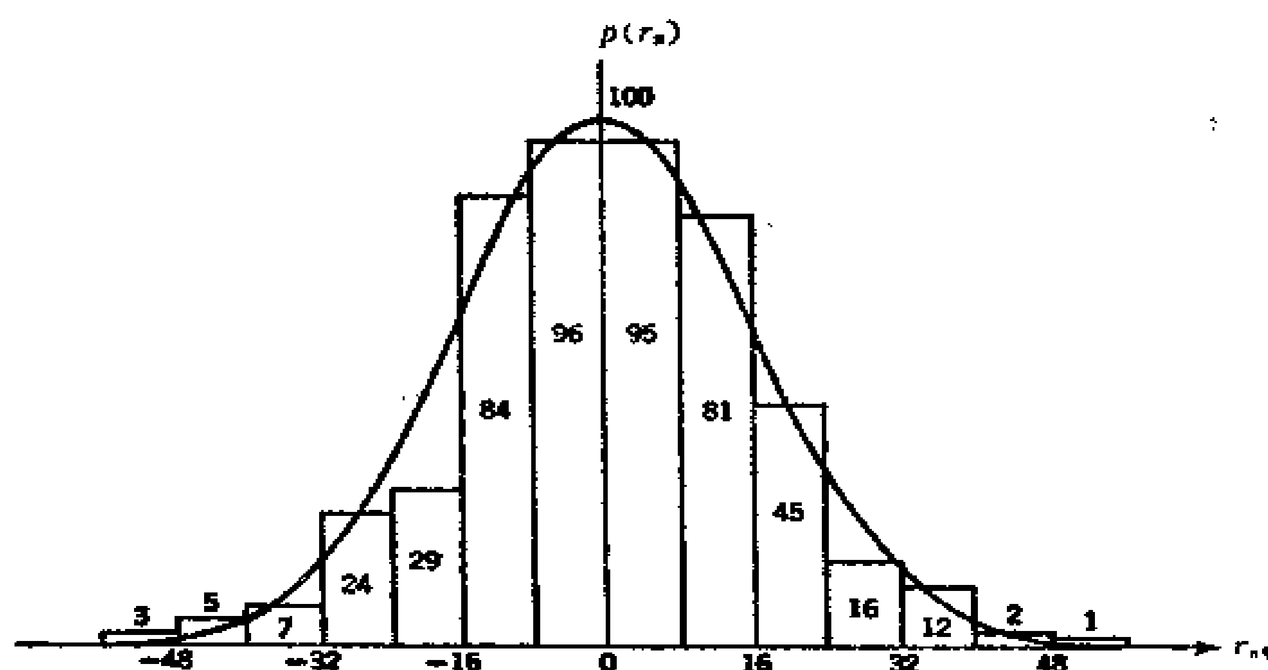


图 1.6 舍入误差的分布

1) 可以得到检验统计理论的假设的精确方法，但它们超出本书范围.

在上面所叙述的相同条件下，所进行的数值试验给出

$$E(r_{1000})_e = -0.14u,$$

$$\operatorname{var}(r_{1000})_c = 39.4u^2.$$

令人满意的是它仍与理论一致。

1.7. 求解的问题

星号表示困难超过一般的问题。

1.2.

1. 证明用 Euler 方法不能近似初值问题

$$y' = y^{1/2},\ y(0) = 0$$

的解 $y = \left(\frac{2}{3}x\right)^{3/2}$，试说明之。

2. 确定以下函数的 Lipschitz 常数：

$$\text{(a)}\ f(x,y) = \frac{2}{x}y \quad (x \geqslant 1);$$

$$\text{(b)}\ f(x,y) = \operatorname{arc\,tg} y;$$

$$\text{(c)}\ f(x,y) = \frac{(x^3-2)^{27}}{17x^2+4}.$$

3*. 利用

$$y_p(x) = y_0 + \int_a^x f_p(t)dt$$

并令 $p \to \infty$ 给出 $y(x) = \lim_{p\to\infty} y_p(x)$ 是微分方程 $y' = f(x,y)$ 的解的另一个证明（指明所需要的全部步骤）。

4*. 给出 $y(x)$ 是 $y' = f(x, y)$ 的唯一解的另一个证明，其方法是注意到任何二个解的差 $\Delta(x)$ 必须满足

$$|\Delta(x)| \leqslant K,$$

其中 K 是常数，$|\Delta(x)| \leqslant L\int_a^x |\Delta(t)|dt$，重复积分，则对于

每一个正整数 n，有

$$|\Delta(x)| \leqslant K \frac{L^n}{n!}(x-a)^n,$$

这仅当 $\Delta(x)=0$ 才有可能.

1.3.

5. 解析地确定初值问题 $y'=2x^{-1}y$，$y(0)=1$ 的 Euler 近似解，同时求出问题的准确解并确定伸缩误差函数。验证关系式(1-49)。[提示：得到量 y_n 的一个简单递推关系式]

6. 用 Euler 方法计算

$$y'=\left(\frac{1}{x}+1\right)y, \quad y(1)=e$$

的近似解的离散误差有多大(近似地)?

7. 求出在定理 1.5 中出现的常数 C_3 的界，用函数 $f(x,y)$ 及其导数的界来表示。

8*. 考虑函数 $y(x,h)$ 的线性差分方程

$$y(x+h,h)-y(x,h)=h[f(x)y(x,h)+g(x)], \tag{1-119}$$

$$y(a,h)=y_0,$$

其中 h 起着参数的作用. 证明以下无穷级数(不一定收敛):

$$y(x,h)=\sum_{p=0}^{\infty} h^p \eta_p(x) \tag{1-120}$$

形式上满足(1-119)，其中

$$\eta_0'=f(x)\eta_0+g(x), \quad \eta_0(a)=\eta_0$$

及

$$\eta_1'=f(x)\eta_1-\frac{1}{2}\eta_0'',$$

$$\eta_2'=f(x)\eta_2-\frac{1}{2}\eta_1''-\frac{1}{6}\eta_0''',$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\eta_k' = f(x)\eta_k - \frac{1}{2}\eta_{k-1}'' - \cdots - \frac{1}{(k+1)!}\eta_0^{(k+1)},$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\eta_k(a) = 0, \quad k = 1, 2, \cdots.$$

9*. 证明：对于

(a) $f(x) = 1,\ g(x) = 0,\ a = 0$;

(b) $f(x) = -\frac{1}{x},\ g(x) = 0,\ a = 1$,

级数（1-120）是收敛的．确定收敛半径并且计算出前几个函数 $\eta_p(x)$．

10．初值问题

$$y' = -2y + \frac{1}{2}x^2 - \frac{1}{3}x^3 - \frac{1}{6}x^4,$$

$$y(0) = \frac{1}{8} \tag{1-121}$$

有准确解

$$y(x) = \frac{1}{8} - \frac{1}{4}x + \frac{1}{4}x^2 - \frac{1}{12}x^4.$$

对步长 $h = 2^{-p}(p = 1, 2, \cdots, 8)$，使用 Euler 方法，得到 $y(1) = \frac{1}{24}$ 的如下近似 $y(1, 2^{-p})$：

p	$y(1, 2^{-p})$
1	0.036458333
2	0.040486653
3	0.041414746
4	0.041608960
5	0.041652882
6	0.041663299
7	0.041665834
8	0.041666459

验证：对于这种情形，从取值来断定不仅 $\lim_{h\to 0} h^{-1}e(1,h)$ 而且 $\lim_{h\to 0} h^{-2}e(1,h)$ 都是存在的。这与定理 1.5 所能预料的相反。同时，验证在这种情形下，外推到极限并不改进收敛性。试用确定伸缩误差函数来说明之。

11*. 利用问题 8 的结果，证明在问题 10 中讨论的近似误差 $e(1,h)$ 满足

$$\lim_{h\to 0} h^{-2}e(1,h) = \frac{5e^{-2}-1}{24} = 0.0135\cdots.$$

12. 解析地确定用 Euler 方法求解初值问题 $y' = x - x^3$, $y(0) = 0$ 所得到的值 y_n. 计算误差 $e_n = y_n - y(x_n)$ 并且把它与由伸缩误差提供的近似值相比较。当 h 很小时，有使误差特别小的 x 值吗(除去 0)?

$$\left[\text{利用公式 } 1+2+\cdots+n = \frac{n(n+1)}{2}\right.$$

$$\left. 1^3+2^3+\cdots+n^3 = \left[\frac{n(n+1)}{2}\right]^2. \right]$$

13*. 已知

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x.$$

如果 K 是一个常数，并且 $|\theta_n| \leqslant 1 (n = 1,2,\cdots)$, 证明

$$\lim_{n\to\infty}\left(1+\frac{x}{n}+\theta_n\frac{K}{n^2}\right)^n = e^x.$$

[利用定理 1.4.]

1.4.

14. 试确定最大容许的局部舍入误差 ε, 如果使用

$$h = 10^{-3},$$

在 1000 步后使问题 $y' = (-1/x)y$, $y(1) = 1$ 的数值积分的累积舍入误差不超过 10^{-4}.

(a) 利用定理 1.6；

(b) 利用定理 1.7.

1.6.

15. 确定乘积 hf 的舍入误差的分布函数 $F(x)$，如果 h 和 f 都是N位二进制数，并且假设取其所有值概率相同. 试确定这个分布的均值与方差.

16. 在定理 1.9 的假设下，并且假设所有乘积 hf 都是舍入到最接近于而不超过 hf 的N位数(去尾舍入). 试确定乘积 hf 的舍入误差分布，并求这个分布的 μ 和 σ^2.

17. 使用十进制数计算并且舍去之. 用 Euler 方法在求解 $y'=y$ 时，对累积舍入误差的分布进行一次小的数值试验. 把这些结果与由定理 1.9 连同问题 16 的结果预估的值相比. [使用 1.6-3 的记号，取 $y_0=1$，$\Delta=0.10101$，$q=0,\cdots,9$，$u=10^{-5}$，$h=0.125$，给出 $E(r_8)_c=-5.6u$，$\mathrm{var}(r_8)_c=1.36u^2$].

18. 利用由 (1-67) 确定的近似值 $d_m(x_n)$ 代入 (1-114)，并且将其当作定积分的近似和，试给出 (1-117) 的另一个证明.

19*. 如果函数 $m(x)$ 和 $v(x)$ 分别由 (1-72) 及 (1-116) 确定，证明：对于 $x>a$，

$$[v(x)]^{1/2}\geqslant\frac{1}{(x-a)^{1/2}}m(x).$$

作为一个应用，证明不等式

$$\left(\frac{e^{2x}-1}{2}\right)^{1/2}\geqslant\frac{1}{x^{1/2}}(e^x-1).$$

[提示：用定积分求解微分方程，并且应用 Schwarz 不等式.]

注

1.1. Euler方法又称为 Euler-Cauchy 折线方法；Cauchy 曾

先在[1840]给出收敛性的精确陈述.

1.2. 这里给出的存在性定理及其证明不同于 Peano 的存在性定理(例如见 Kamke [1947, p.59]; Coddington 及 Levinson [1955, p.6]),它们也是基于折线的方法. 其方法从开始就使用 Lipschitz 条件,这就可能避免利用(非构造性) Ascoli 定理,并且是用整个序列 $\{y_p(x)\}$ 而不是用其子序列来证明其收敛性.

1.3. Collatz [1960, p.58] 的证明与定理 1.3 稍许不同.

1.6. 微分方程逐步积分法舍入误差的传播的统计模型是由 Brouwer [1937] 引入的,后来 Rademacher [1948], Papoulis [1952], Mikulaschkova [1957] 进行了改进. Huskey [1949], Forsythe [1959], van Wijngaarden [1953] 对这个模型提出反对意见(特别是反对独立性的假设). Forsythe [1959] 建议用人工生成随机误差来代替"自然"舍入误差. 用 1.6-3 中指出的方法来考验这个方法是有意义的.

第二章　一阶单个方程的一般单步方法

在第一章已经证明，通过选取充分小的步长 h，Euler 方法可达到任意高的精确度，如果所完成的运算舍入误差可忽略不计的话. 离散误差大致上与 h 成比例，意即在给定的区间上计算函数 $f(x,y)$ 值 N 次可达到的精确度是与 $\frac{1}{N}$ 成比例的. 本章将证明，存在着下述意义下比 Euler 方法有效得多的方法，即 N 次求值，其精确度可达到与 $\frac{1}{N^p}$ 成比例，对 $p>1$.

求解微分方程的任何一个算法称为单步方法，如果仅知道 x_n, y_n 和 h，用这个算法便可计算出解在 x_{n+1} 点的近似值 y_{n+1}. 为了方便，把 y_{n+1} 对量 x_n, y_n, h 的函数依赖关系写成

$$y_{n+1}-y_n=h\Phi(x_n,y_n;h), \tag{2-1}$$

函数 Φ 称为增量函数，它依赖于所给定的微分方程. 例如，在 Euler 方法中，

$$\Phi(x,y;h)=f(x,y), \tag{2-2}$$

此时 Φ 正好与 h 无关.

量 $h\Phi(x_n,y_n;h)$ 表示近似解的增量. 为了便于与精确解相比，我们引入以后所需要的特殊记号. 如果函数

$$f=f(x,y)$$

满足定理 1.1 的条件以及 (x,y) 是区域

$$a\leqslant x\leqslant b,\ -\infty<y<+\infty$$

内任意一点，其中 f 是给定的，那么对于 $t\in[x,b]$，初值问题

$$z'=f(t,z),\quad z(x)=y \tag{2-3}$$

存在一个解．对同一个定理稍作修改（x 换成 $-x$），解 $z(t)$ 便可通过区间 $[a,x]$ 向后延伸，从而所定义的函数 $z(x)$ 在几何上可以看成是通过点 (x,y) 的微分方程 $z'=f(t,z)$ 的解．对于使得 $x+h\in[a,b]$ 的任意 h，我们规定函数 Δ 为

$$\Delta(x,y;h)=\begin{cases}\dfrac{z(x+h)-y}{h}, & h\neq 0,\\ f(x,y), & h=0,\end{cases} \tag{2-4}$$

且称它为精确相对增量．

由于 $z(t)$ 是 (2-3) 的解，故作为 h 函数的 Δ 在 $h=0$ 处是连续的．量 $h\Delta(x,y;h)$ 表示给定的微分方程的精确解在点 (x,y) 上的增量；对于 $h\neq 0$，Δ 本身就代表一个差商．

2.1. 特殊单步方法

2.1-1．直接利用 Taylor 展式．如果函数 $y(x)$ 是

$$y'=f(x,y)$$

的一个解，这里假设 $f(x,y)$ 无限次可微，那么 $y(x)$ 的高阶导数可由函数 $f(x,y)$ 完全确定．例如，

$$\begin{aligned}y''&=\frac{d}{dx}f(x,y)=f_x(x,y)+f_y(x,y)y'\\&=f_x(x,y)+f_y(x,y)f(x,y)\end{aligned} \tag{2-5}$$

以及用类似的方法可得 $y''',y^{\mathrm{IV}},\cdots$．为使记号简化，我们令

$$y''=f'(x,y),\quad y'''=f''(x,y).$$

一般地，

$$y^{(p+1)}=f^{(p)}(x,y)\quad(p=1,2,\cdots).$$

撇表示对 x 的全导数．

利用这个记号，通过点 (x,y) 的解的增量为

$$z(x+h)-z(x)=hf(x,y)+\frac{h^2}{2}f'(x,y)+\cdots$$
$$+\frac{h^p}{p!}f^{(p-1)}(x,y)+O(h^{p+1}).$$

于是，对于精确的相对增量，我们求得展开式为

$$\Delta(x,y;h)=f(x,y)+\frac{h}{2}f'(x,y)+\cdots$$
$$+\frac{h^{p-1}}{p!}f^{(p-1)}(x,y)+O(h^p),\qquad(2\text{-}6)$$

这里 p 是一个任意正整数.

看来用这样的办法选取增量函数 $\Phi(x,y;h)$ 是很自然的，这就是使得 $\Phi(x,y;h)$ 与 Δ 尽可能一致. 达到这个目的一个明显的方法是令

$$\Phi(x,y;h)=f(x,y)+\frac{h}{2}f'(x,y)+\cdots$$
$$+\frac{h^{p-1}}{p!}f^{(p-1)}(x,y).\qquad(2\text{-}7)$$

由 (2-7) 规定的方法称为 p 阶 Taylor 展式方法. 令 $p=1$，又得到 Euler 方法. 方法的精确度随着 p 的增大可任意提高.

Taylor 展式方法不便于计算是明显的，现在要用同时运算 p 个函数 $f, f', \cdots, f^{(p-1)}$ 来代替所给出的一个函数 f. 在自动计算情形，这就增加了贮存代码所需要的空间. 而且，对最简单的方程（通常可用封闭形式积分），函数 $f^{(k)}$ 的解析表达式的复杂性是随着 k 增大而增大. （读者可对看来很简单的方程，例如 $y'=x^2+y^2$ 来检验这个结论.）因此，并不推荐这个方法于实际使用.

2.1-2. *间接利用 Taylor 展式 (Runge-Kutta 型方法).*

这里的基本思想是要产生一个关于 Φ 的公式，而无需计

算 f 的任何阶导数，使 Φ 与 Δ 的误差为 $O(h^p)$. 我们用下面的简单例子来说明这个思想. 试令

$$\Phi(x,y;h) = a_1 f(x,y) + a_2 f(x + p_1 h, y + p_2 h f(x,y)),$$

其中 a_1, a_2, p_1, p_2 都是待定常数. 对两组不同的自变量计算 f 两次便形成这个量. 按 h 的幂次展开，得到

$$\Phi(x,y;h) = (a_1 + a_2) f(x,y) + h a_2 [p_1 f_x(x,y) + p_2 f_y(x,y) f(x,y)] + O(h^2).$$

利用它与量 $\Delta(x,y;h)$ 尽可能一致的条件，由于 (2-5)，$\Delta(x,y;h)$ 可展开如下形式

$$\Delta(x,y;h) = f(x,y) + \frac{1}{2} h [f_x(x,y) + f_y(x,y) f(x,y)] + O(h^2).$$

比较常数项，求得

$$a_1 + a_2 = 1; \tag{2-8a}$$

使 h 的线性项一致就需要

$$a_2 p_1 = \frac{1}{2}, \quad a_2 p_2 = \frac{1}{2}. \tag{2-8b}$$

可以证明，不增加 $f(x, y)$ 的条件便不能得到二次 项一致. 求解方程 (2-8)，得到解族

$$a_1 = 1 - \alpha, \ a_2 = \alpha, \ p_1 = p_2 = \frac{1}{2\alpha},$$

其中 α 是自由参数，$\alpha \neq 0$. 根据上面的推导，对于 $\alpha \neq 0$，增量函数

$$\Phi(x,y;h) = (1 - \alpha) f(x,y) + \alpha f\left(x + \frac{h}{2\alpha},\ y + \frac{h}{2\alpha} f(x,y)\right) \tag{2-9}$$

与 $\Delta(x,y;h)$ 的偏差仅为 $O(h^2)$，而它不含有 $f'(x,y)$. (2-9) 的二个特殊情形是众所周知的:

(i) $\alpha = \frac{1}{2}$. 在这里 (2-1) 形如

$$y_{n+1} = y_n + \frac{1}{2} h[f(x_n, y_n) + f(x_n + h, y_n + hf(x_n, y_n))]. \tag{2-10a}$$

这个方法有时称为"改进的 Euler 方法"(Collatz [1960, p. 54]); 也可称它为 Heun 方法 (Heun [1900]; Rademacher [1948]).

(ii) $\alpha = 1$. 得到公式

$$y_{n+1} = y_n + hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hf(x_n, y_n)\right), \tag{2-10b}$$

称它为改进的折线方法或修改的 Euler 方法 (Collatz [1960, p. 53]).

(2-10) 的两个方法可作简单的图解说明，我们留给读者. 每积分一步，这两个方法都要求计算函数 $f(x,y)$ 值两次. 如果 $f(x, y)$ 确实不依赖于 y，微分方程的解便简化成计算一个积分，(2-10a) 简化成数值积分的梯形法则，而 (2-10b) 则是中点法则.

由 (2-9) 确定的一般方法也称为简化的 Runge-Kutta 方法. 函数 (2-9) 可以看成在不同点上取函数 f 值的加权平均，并且具有第二项的变元依赖于第一项的特点. 经典的 Runge-Kutta 方法 (Runge [1895]; Kutta [1901])——或许是所有单步方法中最著名的——实质是类似的. 这里 Φ 是计算函数值四次加权平均，并且每次计算中的自变量依赖于前面的计算. 具体地,

$$\Phi(x, y; h) = \frac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4], \tag{2-11a}$$

其中

$$
\begin{cases}
k_1 = f(x, y), \\
k_2 = f\left(x + \frac{1}{2}h,\ y + \frac{1}{2}hk_1\right), \\
k_3 = f\left(x + \frac{1}{2}h,\ y + hk_2\right), \\
k_4 = f(x + h,\ y + hk_3).
\end{cases}
\tag{2-11b}
$$

一个单步方法称为 p 阶的，如果它的增量函数以 h^p 阶的误差来近似精确的相对增量 Δ. 整数 p 称为方法的精确阶，如果这个结论对任何较大整数是不成立的. 于是，由(2-7)确定的 Taylor 展式方法除去当 $f^{(k)}(x, y) = 0$，时，$k \geqslant p$，则是一个精确的 p 阶方法.

构造出的简化 Runge-Kutta 方法(2-9)是 2 阶的. 第三章将证明，由(2-11)确定的方法具有 4 阶，并且一般来说这是方法的精确阶. 后一个说明则可从如下的讨论中推知：如果 $f(x,y)$ 不依赖于 y，则 $k_2 = k_3$，并且 Runge-Kutta 方法变成数值积分的 Simpson 法则，在这里所规定的意义下[1]，知其为 4 阶.

计算研究. 用任何一个单步方法求解初值问题的框图，看上去都象 Euler 方法框图(图 1.1)，只不过情况更为复杂，计算 $f(x_n, y_n)$ 的位置由求更为复杂的函数 $\Phi(x_n, y_n; h)$ 值所代替. 我们将讨论对经典 Runge-Kutta 方法(2-11)Φ 的求值. 把(2-11)写成如下形式：

$$
\Phi(x, y; h) = \sum_{\nu=1}^{4} a_\nu k_\nu, \tag{2-12a}
$$

其中 $k_\nu = f(X_\nu, Y_\nu)$，取 $X_1 = x$，$Y_1 = y$ 以及

---

1) 早先 Runge 对方法的研究，其动机是试图把 Simpson 法则推广到微分方程的积分上去。

$$X_\nu = x + hp_\nu, \quad Y_\nu = y + hp_\nu k_{\nu-1}, \quad \nu = 2,3,4, \tag{2-12b}$$

常数 a_ν 及 p_ν 值为

ν	1	2	3	4
a_ν	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$
p_ν	0	$\frac{1}{2}$	$\frac{1}{2}$	1

被一劳永逸地贮存起来． 据此，Φ 的求值本身可以组成一个循环过程．图 2.1 给出这个方法的完整的框图．（在 2.3-5 中

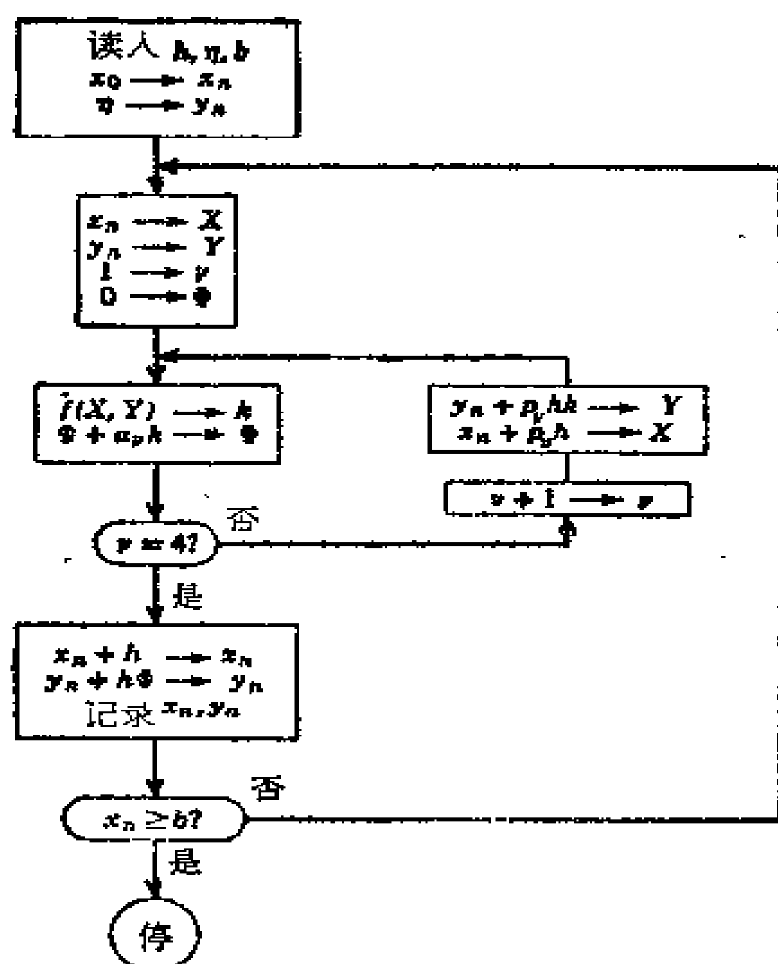


图 2.1　Runge-Kutta 方法框图

表 2.1 Runge-Kutta 方法的数值例子

x_n	y_n	X_ν / Y_ν / $k_\nu = f(X_\nu, Y_\nu)$: $\nu = 1$	2	3	4
0	0	0	0.05000 00000	0.05000 00000	0.10000 00000
		0	0	0.00250 00000	0.00499 93750
		0	0.05000 00000	0.04999 37500	0.09997 50062
		$\Phi = 0.04999\ 37510$			
0.1	0.00499 93751	0.10000 00000	0.15000 00000	0.15000 00000	0.20000 00000
		0.00499 93751	0.00999 81254	0.01249 43770	0.01998 37642
		0.09997 50062	0.14990 00375	0.14984 38905	0.19960 06492
		$\Phi = 0.14984\ 39185$			
0.2	0.01998 37670	0.20000 00000	0.25000 00000	0.25000 00000	0.30000 00000
		0.01998 37670	0.02996 37995	0.03243 88755	0.04487 52950
		0.19960 06491	0.24910 21707	0.24891 52805	0.29798 62034
		$\Phi = 0.24893\ 69551$			
0.3	0.04487 74629				

将讨论关于减少 Runge-Kutta 方法及类似方法的舍人误差的一个特殊方法.)

作为一个数值例子，在表 2.1 中指出用步长 $h=0.1$，Runge-Kutta 方法对初值问题

$$y'=x-y^2,\quad y(0)=0$$

(在 1.1-2 中已经讨论过)的解的前几步. 箭头表示计算这些值的顺序. 由于手算不便于检验，这便成为这个方法的缺点.

2.1-3. 其它单步方法. 除去上面讨论的方法，在文献中还提出许多其它方法. 由于篇幅所限，不详述. 本章末的问题 3—9，12—14 及 18 中讨论一些补充的方法.

2.2. 一般单步方法的离散误差

2.2-1. 收敛性和相容性. 令 $\Phi(x,y;h)$是给定的微分方程 $y'=f(x,y)$ 近似解的单步方法的一个增量函数. 我们将总是假设函数 $f(x,y)$ 满足定理 1.1 的条件，于是初值问题

$$y'=f(x,y),\quad y(a)=\eta \tag{2-13}$$

对于任意 η 有一个解 $y(x)(x\in[a,b])$. 我们还将假设增量函数 $\Phi(x,y;h)$ 对于 $x\in[a,b]$，$y\in(-\infty,+\infty)$ 以及对一切充分小的 $h\geqslant 0$ 在任何给定的点 (x,y) 也是确定的. 从 $y_0=\eta$ 出发，由 (2-1) 只要$x_n\in[a,b]$ 就可能逐个计算 y_n 值.

由增量函数规定的方法称为收敛的，如果对任意 η 和任意 $x\in[a,b]$，

$$\lim_{\substack{h\to 0\\ x_n=x}} y_n=y(x) \tag{2-14}$$

成立. 例如，由定理 1.2，Euler 方法则是一个收敛方法.

关于一般单步方法收敛性的一个简单的必要和充分条件可由以下定理给出.

定理 2.1. 令函数 $\Phi(x,y;h)$ 在由 $x\in[a,b]$, $y\in(-\infty, +\infty)$, $0\leqslant h\leqslant h_0$ 所确定的区域内是连续的（当作它的三个自变量的函数），并且假定存在常数 L，使得对于刚刚定义的区域内的一切 $(x, y; h)$ 和 $(x, y^*; h)$，都有

$$|\Phi(x,y^*;h)-\Phi(x,y;h)|\leqslant L|y^*-y| \quad (2\text{-}15)$$

成立．则关系式

$$\Phi(x,y;0)=f(x,y) \quad (2\text{-}16)$$

便是由增量函数 Φ 所确定的方法的收敛性的必要和充分条件．

条件 (2-16) 称为相容性条件．粗略地说，定理指出相容性是收敛的必要和充分条件．读者容易确信在 2.1 中讨论的所有特殊方法都是相容的．

证．令 $\Phi(x, y; 0)=g(x, y)$，则函数 $g(x,y)$ 满足定理 1.1 的条件．从而推出，对于任意 η，初值问题

$$z'=g(x,z), \quad z(a)=\eta \quad (2\text{-}17)$$

有唯一解 $z(x)$．而且，从定理 1.1 的证明中得出，对于 $x\in[a, b]$，点 $(x, z(x))\in R$，其中 R 表示由 1.2-4 中确定的紧致区域．

我们将证明由 $z_0=\eta$ 及

$$z_{n+1}=z_n+h\Phi(x_n,z_n;h), \quad n=0,1,2,\cdots;x_n\in[a,b] \quad (2\text{-}18)$$

所确定的数 z_n 收敛于 $z(x)$．由 (2-18) 减去关系式

$$z(x_{n+1})=z(x_n)+h\Delta(x_n,z(x_n);h),$$

我们求得关于误差 $e_n=z_n-z(x_n)$ 的关系式

$$e_{n+1}=e_n+h[\Phi(x_n,z_n;h)-\Delta(x_n,z(x_n);h)].$$

利用中值定理，

$$\Delta(x_n,z(x_n);h)=\frac{z(x_{n+1})-z(x_n)}{h}$$
$$=g(x_n+\theta h,z(x_n+\theta h)),$$

其中 $0<\theta<1$，括号内的表达式从而可写成

$$\begin{aligned}&\Phi(x_n,z_n;h)-\Phi(x_n,z(x_n);h)\\&\quad+\Phi(x_n,z(x_n);h)-\Phi(x_n,z(x_n);0)\\&\quad+g(x_n,z(x_n))-g(x_n+\theta h,z(x_n+\theta h)).\end{aligned}$$

连续函数 $\Phi(x,y;h)$ 在紧致集

$$x\in[a,b],\ y=z(x),\ 0\leqslant h\leqslant h_0$$

上是一致连续的．因此象在 1.2-4 中一样，我们可以断定量

$$\zeta(h)=\max_{x\in[a,b]}|\Phi(x,z(x);h)-\Phi(x,z(x);0)|$$

当 $h\to 0$ 时趋于零．类似地，

$$\chi(h)=\max_{\substack{x\in[a,b]\\0\leqslant k\leqslant h}}|g(x,z(x))-g(x+k,z(x+k))|$$

当 $h\to 0$ 时趋于零．利用 (2-15)，于是我们得到估计

$$|e_{n+1}|\leqslant|e_n|+h[L|e_n|+\zeta(h)+\chi(h)].$$

这是形式 (1-11)，而取 $A=1+hL$，$B=h[\zeta(h)+\chi(h)]$。于是导出

$$|e_n|\leqslant[\zeta(h)+\chi(h)]E_L(x_n-a).$$

由于右端的表达式当 $h\to 0$ 时趋向 0，从而便有

$$\lim_{\substack{h\to 0\\x_n=x}}z_n=z(x),\ x\in[a,b].$$

这就证明了相容性条件 (2-16) 对收敛性是充分的．如果 $g(x,y)=f(x,y)$，那么 $z(x)=y(x)$．为了证明必要性，假设由 $\Phi(x,y;h)$ 确定的方法是收敛的，但在某个点 (x,y) 上，$g(x,y)\neq f(x,y)$．从定理 1.1 推出，存在一个 η 使得初值问题的解 $y(t)$ 通过点 (x,y)．$z_0=\eta$ 和 (2-18) 确定的值 z_n 一方面趋向 $y(t)$；另一方面，利用上面的证明，z_n 又趋向初值问题 (2-17) 的解 $z(t)$．如果 $z(x)\neq y(x)$，我们立即得到矛盾；如果 $z(x)=y(x)$，那么

$$z'(x)=g(x,y)\neq f(x,y)=y'(x),$$

仍得 $z(t) \equiv y(t)$。 于是 $g(x, y) \equiv f(x, y)$ 的假设便导致矛盾.

2.2-2. 一个先验界. 虽然定理 2.1 保证了一个相容方法为收敛的,但并不能用它来估计离散误差. 特别是,它没有指出(累积)离散误差的阶. 以下的结果弥补了这个缺陷. 我们用 $y(x)$ 表示初值问题 (2-13) 的解, 并令 Δ 是由 (2-4) 确定的. 于是有

定理 2.2. 令 $\Phi(x,y;h)$ 满足定理 2.1 的条件, 并且存在常数 $N \geqslant 0$ 和 $p \geqslant 0$, 使得

$$|\Phi(x,y(x);h) - \Delta(x,y(x);h)| \leqslant Nh^p, \quad (2\text{-}19)$$
$$x \in [a,b], \ h \leqslant h_0,$$

那么,对于 $x_n \in [a,b]$ 和任意 $h \leqslant h_0$, 有

$$|y_n - y(x_n)| \leqslant h^p N E_L(x_n - a), \quad (2\text{-}20)$$

其中 E_L 表示 Lipschitz 函数.

出现在 (2-19) 中的量 $\Phi - \Delta$ 称为方法的相对局部离散(或截断)误差[1]. 这个定理可粗略地叙述为,如果相对局部离散误差为 $O(h^p)$, 那么(绝对)累积离散误差也有相同的阶.

证. 从 (2-1) 减去恒等式 $y(x_{n+1}) = y(x_n) + h\Delta(x_n, y_n;h)$, 并令 $e_n = y_n - y(x_n)$, 把这个结果写成形式

$$e_{n+1} = e_n + h[\Phi(x_n,y_n;h) - \Phi(x_n,y(x_n);h) + \Phi(x_n,y(x_n);h) - \Delta(x_n,y(x_n);h)]. \quad (2\text{-}21)$$

利用 (2-15) 及 (2-19) 来估计右端的表达式,我们求得

$$|e_{n+1}| \leqslant |e_n| + hL|e_n| + h^{p+1}N.$$

这个不等式形式为 $|e_{n+1}| \leqslant A|e_n| + B$, 其中 $A = 1 + hL$, $B = h^{p+1}N$. 应用引理 1.2 便推出所需要的结果.

象定理 1.4 与定理 1.3 有关一样,按照同样的方式, 以下

1) 相对是指对步长 h 而言.

结果与定理 2.2 有关。

定理 2.3. 令函数 $\Phi(x,y;h)$ 满足定理 2.2 的条件，并令 $\{y_n\}$ 是满足

$$y_0 = \eta; \tag{2-22}$$

$$y_{n+1} = y_n + h[\Phi(x_n,y_n;h) + h^q\theta_n K],$$

$$n = 0,1,2,\cdots;\ x_n \in [a,b]$$

的任何一个数列，其中 $K \geqslant 0$ 和 $q \geqslant 0$ 都是常数，并且 θ_n 是使得 $|\theta_n| \leqslant 1$ 的任意数，那么，对于 $x_n \in [a,b]$ 和 $h \leqslant h_0$，

$$|y_n - y(x_n)| \leqslant h^r N_1 E_L(x_n - a), \tag{2-23}$$

其中 $r = \min(p,q)$ 和 $N_1 = Nh_0^{p-r} + Kh_0^{q-r}$。

这个定理保证了值 y_n 收敛于精确解，即使基本关系式 (2-1) 仅是近似地满足，如果对 $q > 0$ 每一步的误差为 $O(h^{1+q})$。

定理证明是定理 2.2 的证明的直截了当的修改，把常数 N 全部换成 N_1。

2.2-3. *特殊方法的应用.* L 的表达式. 为了具体应用定理 2.2 的误差界，必须知道分别出现在 (2-15) 和 (2-19) 中的常数 L 和 N 的值。在本节和下节中，我们将对 2.1 中所讨论的特殊方法来计算这些值。

假设函数 $f(x,y)$ 是连续的且有充分高阶的连续导数。我们令

$$L_k = \sup_{\substack{x\in[a,b]\\ y\in(-\infty,+\infty)}} \left|\frac{\partial}{\partial y} f^{(k)}(x,y)\right|,\quad k = 0,1,\cdots,$$

并且假设这些量都是有限的。于是常数 L_0 可作为函数 $f(x,y)$ 本身的 Lipschitz 常数。

对于基于 Taylor 展式的方法 (2-7)，应用中值定理，如果 $h \leqslant h_0$，我们求得

$$L \leqslant L_0 + \frac{h_0}{2} L_1 + \cdots + \frac{h_0^{p-1}}{p!} L_{p-1}.$$

对于简化的 Runge-Kutta 方法(2-9)，结合不等式

$$|f(x,y^*)-f(x,y)|\leqslant L_0|y^*-y|$$

及

$$\left|f\left(x+\frac{h}{2\alpha},y^*+\frac{h}{2\alpha}f(x,y^*)\right)-f\left(x+\frac{h}{2\alpha},y+\frac{h}{2\alpha}f(x,y)\right)\right|$$

$$\leqslant L_0\left|y^*-y+\frac{h}{2\alpha}[f(x,y^*)-f(x,y)]\right|$$

$$\leqslant L_0\left(1+\frac{h}{2|\alpha|}L_0\right)|y^*-y|,$$

我们求得常数 L 界为

$$L\leqslant|1-\alpha|L_0+|\alpha|\left(1+\frac{h_0L_0}{2|\alpha|}\right)L_0.$$

如果 $0<\alpha\leqslant 1$，便导出

$$L\leqslant\left(1+\frac{h_0L_0}{2}\right)L_0. \tag{2-24}$$

可类似地计算出经典 Runge-Kutta 方法的 L 常数为(Carr [1958])

$$L\leqslant\left(1+\frac{h_0L_0}{2}+\frac{h_0^2L_0^2}{6}+\frac{h_0^3L_0^3}{24}\right)L_0\leqslant\frac{e^{h_0L_0}-1}{h_0}. \tag{2-25}$$

2.2-4. 特殊方法的应用． N的表达式． 如果精确解 $y(x)$为未知，从(2-19)来估计N当然是不可能的．但是从定理 1.1 的证明，我们知道精确解 $y(x)$ 位于由 1.2-4 中确定的紧致区域R内．因此，对于 $h\leqslant h_0$ 和 $(x,y)\in R$，便可用

$$h^{-p}|\Phi(x,y;h)-\Delta(x,y;h)|$$

的上界来代替 N．如果由 Φ 确定的方法为 p 阶，并且 Φ 和 Δ 都是 p 次连续可微，那么，对于 $k=0,1,\cdots,p-1$，

$$\frac{\partial^k}{\partial h^k}\{\Phi(x,y;h)-\Delta(x,y;h)\}_{h=0}=0,$$

并且从 Taylor 公式[1]，我们得到

$$\Phi(x,y;h)-\Delta(x,y;h)$$
$$=h^p\left\{\frac{1}{p!}\frac{\partial^p}{\partial h^p}\Phi(x,y;h^*)\right.$$
$$\left.-\frac{1}{(p+1)!}f^{(p)}(x+h^*,z(x+h^*))\right\},$$

其中 $0<h^*<h$. 于是 (2-19) 满足，取

$$N=\max_{\substack{(x,y)\in R\\ h\leq h_0}}\left|\frac{1}{p!}\frac{\partial^p\Phi}{\partial h^p}(x,y;h)\right.$$
$$\left.-\frac{1}{(p+1)!}f^{(p)}(x+h,z(x+h))\right|. \quad (2\text{-}26)$$

显然可把这个结果应用于显式 Taylor 展式方法，由于

$$\frac{\partial^p\Phi}{\partial h^p}(x,y;h)=0,$$

我们可简便地求得

$$N=\frac{1}{(p+1)!}\max_{(x,y)\in R}|f^{(p)}(x,y)|.$$

对于简化的 Runge-Kutta 方法 (2-9)，有 $p=2$. 由于

$$\Phi(x,y;h)=(1-\alpha)f(x,y)+\alpha f(x',y'),$$

其中

$$x'=x+\frac{h}{2\alpha},\quad y'=y+\frac{h}{2\alpha}f(x,y),$$

我们求得

$$\frac{\partial\Phi}{\partial h}(x,y;h)=\frac{1}{2}\{f_x(x',y')+f_y(x',y')f(x,y)\}$$

及

$$\frac{\partial^2\Phi}{\partial h^2}(x,y;h)=\frac{1}{4\alpha}\{f_{xx}(x',y')+2f_{xy}(x',y')f(x,y)$$
$$+f_{yy}(x',y')[f(x,y)]^2\}. \quad (2\text{-}27)$$

---

1) $z(t)$ 表示通过 (x,y) 的 $z'=f(t,z)$ 的解.

如果我们确定常数M和K，使得对于$(x,y)\in R$，$M\geqslant 1$，

$$|f(x,y)|\leqslant M,\ \left|\frac{\partial^{i+k}f}{\partial x^i\partial y^k}\right|\leqslant\frac{K}{M^{i+k-1}}\quad(i+k\leqslant 2),\tag{2-28}$$

那么表达式(2-27)可估计如下：

$$\left|\frac{\partial^2\Phi}{\partial h^2}(x,y;h)\right|\leqslant\frac{1}{|\alpha|}KM.$$

从

$$f''=f_{xx}+2f_{xy}f+f_{yy}f^2+f_y(f_x+f_yf)$$

容易求得

$$|f''|\leqslant 4KM+2K^2M.$$

于是对于简化的 Runge-Kutta 方法，

$$N\leqslant KM\left(\frac{1}{2|\alpha|}+\frac{2}{3}+\frac{1}{3}K\right).\tag{2-29}$$

对于经典 Runge-Kutta 方法，Bieberbach (1944, p. 55) 用类似的计算求出简单的界为

$$N\leqslant 6KM(1+K+K^2+K^3+K^4).\tag{2-30}$$

在这里必须假设(2-28)的第二个界对于$i+k\leqslant 4$成立。对于简化的和经典的 Runge-Kutta 这两个方法的N的更好界可以用精确的解析方法来得到(见 3.3-4)。

2.2-5. **主误差函数**. 如果由$\Phi(x,y;h)$确定的方法有精确p阶，并且$\Phi(x,y;h)$和$\Delta(x,y;h)$有关于h的$p+1$阶连续导数，它也依赖于(x,y)，那么我们可写成

$$\Phi(x,y;h)-\Delta(x,y;h)=h^p\varphi(x,y)+O(h^{p+1}),\tag{2-31}$$

其中函数$\varphi(x,y)$连续且恒不为零. 函数$\varphi(x,y)$称为方法的主误差函数.

量$h[\Phi(x,y;h)-\Delta(x,y;h)]$表示从点$(x,y)$出发一步

后的方法的误差．称这个量为方法的(绝对)局部离散误差或局部截断误差．方法的阶 p 可看成对方法的精确度的一个粗糙的量度，因为它说明局部离散误差相当于 h^{p+1}，而不是对 h 的任意高的幂次．主误差函数能使我们给出如下的结论：对于充分小的 h 值，局部离散误差可近似地表示成 $h^{p+1}\varphi(x,y)$.

我们对于在 2.1 中讨论的特殊方法来确定主误差函数．

对于 Euler 方法，我们有

$$\begin{aligned}&\Phi(x,y;h)-\Delta(x,y;h)\\&=f(x,y)-\left[f(x,y)+\frac{1}{2}hf'(x,y)+O(h^2)\right].\end{aligned}$$

如果 $f'(x,y)\neq 0$，方法的精确阶为 1，主误差函数为

$$\varphi(x,y)=-\frac{1}{2}f'(x,y). \tag{2-32}$$

类似地，对于 p 阶 Taylor 展式方法，我们求得

$$\varphi(x,y)=\frac{-1}{(p+1)!}f^{(p)}(x,y). \tag{2-33}$$

对于 Runge-Kutta 型方法，计算并不简便．对于简化的 Runge-Kutta 方法，我们把 (2-9) 的右端表达式扩展到含有 h^2 的项．应用两个自变量函数的 Taylor 公式，求得

$$\begin{aligned}\Phi(x,y;h)=f+\frac{h}{2}(f_x+f_yf)+\frac{h^2}{8\alpha}(f_{xx}+2f_{xy}f+f_{yy}f^2)\\+O(h^3),\end{aligned}$$

把变量 (x,y) 代入到函数 f 和它的导数中．另一方面，写出 (2-6) 中的导数 f' 和 f''，我们有

$$\begin{aligned}\Delta(x,y;h)=f+\frac{h}{2}(f_x+f_yf)\\+\frac{h^2}{6}(f_{xx}+2f_{xy}f+f_{yy}f^2+f_xf_y+f_y^2f)+O(h^3).\end{aligned}$$

对照最后两个关系式，导出

$$\varphi(x,y) = \left(\frac{1}{8\alpha} - \frac{1}{6}\right)(f_{xx} + 2f_{xy}f + f_{yy}f^2)$$

$$- \frac{1}{6}(f_x f_y + f_y^2 f), \qquad (2\text{-}34)$$

这可写成如下的更紧凑的形式:

$$\varphi(x,y) = \left(\frac{1}{8\alpha} - \frac{1}{6}\right)f'' - \frac{1}{8\alpha}f_y f'. \qquad (2\text{-}35)$$

对于经典的 Runge-Kutta 方法,在3.3-5中将作为更一般的情形来讨论的一个附带结果,我们得到主误差函数

$$\varphi(x,y) = \frac{1}{2880}f^{(4)} - \frac{1}{576}f_y f''' + \frac{1}{288}(f_y^2 - f_{xy} - f_{yy}f)f''$$

$$+ \frac{1}{192}(2f_{xy}f_y + 3f_{yy}f_y f - 2f_y^3 + f_{yy}f_x)f', \qquad (3\text{-}36)$$

各处将自变量理解为 (x,y).

例. 在表 2.2 中, 我们对于已讨论过的一些方法列出对应于方程 $y' = y$ 的主误差函数. 对于这个特殊方程, 给出 $\varphi(x,y) = Ky$, 其中常数 K 值依赖于方法.

表 2.2

方法	阶	K
Taylor 展式 (2-7)	p	$\frac{-1}{(p+1)!}$
简化的 Runge-Kutta 方法 (2-9)	2	$\frac{-1}{6}$
经典的 Runge-Kutta 方法 (2-11)	4	$-\frac{1}{120}$

2.2-6. *误差的渐近公式*. 借助于数值例子, 容易证实 2.2-2 中导出的误差估计一般说来大部分都超过真实误差, 尤其是步数较大时. 象第一章一样, 我们宁可用真实误差近似公式而不用其界来补充这些结果.

我们假设由 $\Phi(x, y; h)$ 定义的方法的阶至少为 1, 以及

$f(x,y)$ 和 $\Phi(x,y;h)$ 在区域 $x\in[a,b]$, $y\in(-\infty,\infty)$, $h\leqslant h_0$ 内对阶 $\leqslant p+1$ 的一切偏导数都存在连续且有界，那么在 (2-31) 中由 $O(h^{p+1})$ 表示的项形如 θKh^{p+1}，其中 K 是常数，$|\theta|\leqslant 1$. 我们再一次把由 (2-1) 确定的近似值 y_n 与对应的精确值 $y(x_n)$ 进行比较．由于满足定理 2.2 的假设，我们知道存在一个常数 K_1，使得对于 $x_n\in[a,b]$，$h\leqslant h_0$，

$$|y_n-y(x_n)|\leqslant K_1h^p. \tag{2-37}$$

对于这些信息，我们回到关系式 (2-21)．利用 $y_n=y(x_n)+e_n$ 且展成 e_n 的幂次，得

$$\begin{aligned}
&\Phi(x_n,y_n;h)-\Phi(x_n,y(x_n);h)\\
&\quad=\Phi_y(x_n,y(x_n);h)e_n+\frac{1}{2}\Phi_{yy}(x_n,y^*;h)e_n^2\\
&\quad=[\Phi_y(x_n,y(x_n);0)+h\Phi_{yh}(x_n,y(x_n);h^+)]e_n\\
&\qquad+\frac{1}{2}\Phi_{yy}(x_n,y^*;h)e_n^2,
\end{aligned} \tag{2-38}$$

这里 y^* 是在 $y(x_n)$ 与 y_n 之间的值，$0<h^+<h_0$. 由于 $p\geqslant 1$，由 (2-19) 导出，$\Phi(x,y;0)=f(x,y)$. 我们始终令

$$g(x)=f_y(x,y(x)),$$

因此

$$\Phi_y(x_n,y(x_n);0)=g(x_n).$$

此外，$|e_n|\leqslant K_1h^p$，并且 Φ 的二阶导数有界．利用新的常数 K_3，从而可以写成

$$\Phi(x_n,y_n;h)-\Phi(x_n,y(x_n);h)=g(x_n)e_n+\theta K_3h^{p+1},$$

今后 $\theta,\theta',\cdots$ 表示随方程而改变且模小于 1 的未定数.

代入 (2-21) 且利用 (2-31)，我们得到

$$\begin{aligned}
e_{n+1}=e_n+h\{&g(x_n)e_n+\theta K_3h^{p+1}\\
&+\varphi(x_n,y(x_n))h^p+\theta'Kh^{p+1}\}.
\end{aligned} \tag{2-39}$$

利用关系式

$$e_n=\bar{e}_nh^p,$$

我们引入新的数 $\bar{e}_n$，称它为伸缩误差．由此，关系式 (2-39) 可写成形式

$$\bar{e}_{n+1} = \bar{e}_n + h\{g(x_n)\bar{e}_n + \varphi(x_n, y(x_n))\} \\ + \theta'' K_4 h^2, \tag{2-40}$$

其中 $K_4 = K + K_3$. 把定理 1.4 应用到关系式 (2-40)，于是存在一个常数 K_5，使得

$$\bar{e}_n = e(x_n) + \theta K_5 h,$$

其中函数 $e(x)$ 是初值问题

$$e'(x) = g(x)e(x) + \varphi(x, y(x)), \qquad (2\text{-}41)$$
$$e(a) = 0$$

的解．函数 $e(x)$ 称为伸缩误差函数．

我们已经证明：

定理 2.4. 令 p 是由 $\Phi(x, y; h)$ 确定的方法的精确阶，并且 $\Phi(x, y; h)$ 和 $f(x, y)$ 满足 2.2-6 开始时叙述的条件，那么由 (2-1) 确定的近似解的离散误差满足

$$e_n = h^p e(x_n) + O(h^{p+1}), \tag{2-42}$$

其中 $e(x)$ 表示由 (2-41) 确定的伸缩误差函数．

对于 $f(x, y) = y$，对已经讨论过的所有特殊方法求得 $\varphi(x, y) = Ky$，于是对于初值问题 $y' = y$，$y(0) = 1$，$e(x)$ 的初值问题化成

$$e'(x) = e(x) + Ke^x, \quad e(0) = 0.$$

积分，得到

$$e(x) = Kxe^x.$$

因此，对于表 2.2 中列出的方法，

$$e_n = Kh^p x_n e^{x_n} + O(h^{p+1}).$$

虽然绝对误差随 x 按指数增长，而相对误差却仅是线性增长．

2.2-7. 渐近公式的应用． 在 1.3-4 中对于 Euler方 法所

讨论的 Richardson 外推到 $h=0$ 的一般单步方法，最直接应用定理 2.4 是合适的。为了明确起见，我们用 $y(x;h)$ 表示用步长 h 在点 x（假设 $x-a$ 是 h 的整数倍）所得到的近似值，从而方程（2-42）可重新写成

$$y(x,h)=y(x)+h^{p}e(x)+O(h^{p+1}). \qquad (2-43)$$

对于两个不同的 h 值，比如值 h 和 qh，其中 q 既不是 0 也不是 1，求解微分方程，并且写出对应于（2-43）的方程，我们可以把结果看成为两个未知量 $y(x)$ 和 $e(x)$ 的两个线性方程组。解出 $y(x)$，便得到

$$y(x)=\frac{y(x,qh)-q^{p}y(x,h)}{1-q^{p}}+O(h^{p+1}). \qquad (2-44)$$

这个关系式表示“外推”值（在（2-44）中以分数表示）近似精确解，其误差的阶超过方法的阶为 1．由（2-44），显然 q 值不能接近于 1．在实际中常常采用 $q=2$，而且以上的分析是假设没有舍入误差．如果舍入误差比得上离散误差，便要影响计算值 $y(x,h)$，那么这个外推值就表现出不稳定．

为了得到 e_n 的一个近似值，使用定理 2.4 的另一个方法如下：把 $e(x)$ 的微分方程（2-41）与给定的微分方程一起进行数值积分，并且略去（2-42）的修正项来计算 e_n．如果实现这个方案，看来必须知道：

（a）方法的主误差函数 $\varphi(x,y)$；

（b）导数 $f_y(x,y)$；

（c）问题的真解，因为它用数字表示 f_y 和 φ 中的自变量．

最后的要求乍一看似乎是荒谬的，它是这三个要求中最不重要的．如果以近似值 y_n 代替真解 $y(x)$ 而代入到（2-41）的右端，从而解 $e(x)$ 具有量级为 $O(h^{p})$ 的误差，这可把它合并到（2-42）的误差项 $O(h^{p+1})$ 中．

上面的方法实际上是希望不大的，尤其是因为它还包含

着计算函数 $f_y(x,y)$ 和 $\varphi(x,y)$. 由于这个原因，用单步方法解初值问题的很多现存的方法都是避开对伸缩误差方程的积分. 如果任何试图要去估计离散误差，通常则是控制对局部离散误差的近似计算. 如果局部误差超过预先指定的界限值，则减小积分步长；如果它是比允许的界限小得多，则增加步长. 显然这个方法至多达到对误差定性的控制，用这个方式不可能对累积离散误差有可靠的数量上的估计.

为了近似计算出局部误差，现在介绍 Gorn 和 Moore [1953] 导出的一个简单算法.

令 $z(t)$ 仍表示通过点 (x,y) 的方程 $z'=f(t,z)$ 的解，我们从 (x,y) 出发，用 $\Phi(x,y;h)$ 所确定的方法，或用步长 h 积分一步或用步长 $\frac{1}{2}h$ 积分二步，可以得到 $z(x+h)$ 的一个近似值. 我们用 δ_1 与 δ_2 分别表示其增量，并且把它们两个与精确增量 $h\Delta(x,y;h)$ 进行比较.

如果方法具有 p 阶，对于一步增量，我们得到

$$\begin{aligned}\delta_1 &= h\Phi(x,y;h)\\ &= h\Delta(x,y;h)+h^{p+1}\varphi(x,y)+O(h^{p+2});\end{aligned}\tag{2-45}$$

对于二步增量，给出

$$\begin{aligned}\delta_2 = &\frac{1}{2}h\Phi\left(x,y;\frac{1}{2}h\right)\\ &+\frac{1}{2}h\Phi\left(x+\frac{1}{2}h,y+\frac{1}{2}h\Phi\left(x,y;\frac{h}{2}\right);\frac{h}{2}\right).\end{aligned}$$

由于

$$\begin{aligned}&\frac{1}{2}h\Phi\left(x,y;\frac{1}{2}h\right)\\ &\quad=\frac{1}{2}h\Delta\left(x,y;\frac{1}{2}h\right)+\left(\frac{1}{2}h\right)^{p+1}\varphi(x,y)+O(h^{p+2}),\end{aligned}$$

我们有

$$\delta_2 = \frac{1}{2}h\Delta\left(x,y;\frac{1}{2}h\right) + \left(\frac{1}{2}h\right)^{p+1}\varphi(x,y) + O(h^{p+2})$$
$$+ \frac{1}{2}h\Phi\left(x+\frac{1}{2}h, z\left(x+\frac{1}{2}h\right); \frac{1}{2}h\right)$$
$$+ \frac{1}{2}h\Phi_y\left(x+\frac{1}{2}h, z^*; \frac{1}{2}h\right)$$
$$\times\left[\left(\frac{1}{2}h\right)^{p+1}\varphi(x,y) + O(h^{p+2})\right], \tag{2-46}$$

其中 z^* 是位于 $z\left(x+\frac{1}{2}h\right)$ 和 $y+\frac{1}{2}h\Phi\left(x,y;\frac{1}{2}h\right)$ 之间的一个值．在 δ_2 的表达式中，第三行为 $O(h^{p+2})$，第二行可换成

$$\frac{1}{2}h\Delta\left(x+\frac{1}{2}h, z\left(x+\frac{1}{2}h\right); \frac{h}{2}\right)$$
$$+\left(\frac{1}{2}h\right)^{p+1}\varphi\left(x+\frac{1}{2}h, z\left(x+\frac{1}{2}h\right)\right) + O(h^{p+2}).$$

由于

$$\frac{1}{2}h\Delta\left(x,y;\frac{1}{2}h\right) + \frac{1}{2}h\Delta\left(x+\frac{1}{2}h, z\left(x+\frac{1}{2}h\right); \frac{1}{2}h\right)$$
$$= h\Delta(x,y;h)$$

及

$$\varphi\left(x+\frac{1}{2}h, z\left(x+\frac{1}{2}h\right)\right) = \varphi(x,y) + O(h),$$

于是导出

$$\delta_2 = h\Delta(x,y;h) + 2\left(\frac{1}{2}h\right)^{p+1}\varphi(x,y)$$
$$+ O(h^{p+2}). \tag{2-47}$$

我们规定量

$$\delta = \delta(x,y;h) = \delta_2 - \delta_1. \tag{2-48}$$

对于给定的 x,y 和 h，用步长为 h 一步和步长为 $\frac{1}{2}h$ 二步计算出的结果的差直接计算 δ 值。对照 (2-45) 与 (2-47)，我们求得

$$\delta = -(1-2^{-p})h^{p+1}\varphi(x,y)+O(h^{p+2}). \quad (2\text{-}49)$$

解出 $\varphi(x,y)$，得出主误差函数所需要的近似表达式．这个结果可综合成：

定理 2.5. 在定理 2.4 的条件下，主误差函数满足

$$\varphi(x,y) = -\frac{1}{1-2^{-p}}h^{-p-1}\delta(x,y;h)+O(h), \quad (2\text{-}50)$$

其中 $\delta(x,y;h)$ 是由 (2-48) 直接计算的量．按 δ 的定义推导的算法造诣浅而 Richardson 外推极限的算法造诣深。因此可称之为局部外推极限．

2.2-8．变步长． 对于某些微分方程，尤其是由物理问题引起的那些方程，用定性方法能预估出其解在独立变量某些范围内将有十分光滑的性态，而在其它范围则几乎没有光滑性．星际导弹轨道便是一个例子，除去在天体的附近以外，它是光滑的．在这样情形，对独立变量的整个范围用同一个 h 值来进行积分是不合理的． 我们将修改定理 2.2 和 2.3 的结论来处理这种情形．

假设真实使用的步长为 $\theta(x)h$，其中基本步长 h 是常数，并且 $\theta(x)$ 是使 $0<\theta(x)\leqslant 1$，$x\in[a,b]$ 的 x 的分段连续函数．我们重新规定节点 x_n 为

$$x_0 = a, x_{n+1} = x_n + \theta(x_n)h, \quad n = 0,1,2,\cdots. \quad (2\text{-}51)$$

由 $\theta(x)$ 的假设推出，$\theta(x)$ 有大于零的界，即存在一个常数 $c>0$，使得 $\theta(x)\geqslant c$，$a\leqslant x\leqslant b$ 成立．依次推出

$$x_n \geqslant a + nch;$$

因此对于每一个 $h>0$，在区间 $[a,b]$ 上的积分仅需要有限

步.

近似值 y_n 是由公式

$$y_{n+1}=y_n+\theta(x_n)h\Phi(x_n,y_n;\theta(x_n)h) \qquad (2\text{-}52)$$

生成的. 我们有类似于定理 2.2 的如下定理:

定理 2.6. 令位于 (2-19) 中的函数 $\Phi(x, y; h)$ 满足条件

$$|\Phi(x,y(x);h\theta(x))-\Delta(x,y(x);h\theta(x))|\leqslant Nh^p, \qquad (2\text{-}53)$$

则由 (2-52) 确定的值 y_n 仍满足

$$|y_n-y(x_n)|\leqslant NE_L(x_n-a)h^p,\ x_n\in[a,b]. \qquad (2\text{-}54)$$

证明完全类似于定理 2.2, 故从略. 给出类似于定理 2.4 的定理:

定理 2.7. 令 $\Phi(x,y;h)$ 和 $f(x,y)$ 满足定理 2.4 的条件, 则由 (2-52) 确定的近似解 y_n 误差 e_n 满足

$$e_n=e(x_n)h^p+O(h^{p+1}),\ x_n\in[a,b], \qquad (2\text{-}55)$$

其中函数 $e(x)$ 是由

$$e(a)=0,$$

$$e'(x)=g(x)e(x)+[\theta(x)]^p\varphi(x,y(x)) \qquad (2\text{-}56)$$

确定的. 证明只须把定理 2.4 的证明作一些修改, 因此留给读者去做.

2.2-9. 线性方程; 数值例子. 如果要积分的微分方程形如

$$y'=g(x)y, \qquad (2\text{-}57)$$

即线性齐次[1], 则前几节的一些分析便可简化. 如果增量函数 $\Phi(x,y;h)$ 保持线性(对本章讨论的所有方法都成立), 那么主

---

1) 虽然此种微分方程可用求积来积分, 但由于数值的目的, 可用上面方法之一来直接积分, 因为这可避免计算指数函数值.

误差函数关于 y 也是线性的．我们可令

$$\varphi(x,y)=\phi(x)y, \tag{2-58}$$

则伸缩误差函数 $e(x)$ 满足

$$e'(x)=g(x)e(x)+[\theta(x)]^p\phi(x)y(x). \tag{2-59}$$

除去 $y(x)\equiv 0$ 的平凡情形，我们规定相对伸缩误差函数 $r(x)$ 为

$$r(x)=\frac{e(x)}{y(x)}. \tag{2-60}$$

把 $e(x)=r(x)y(x)$ 代入到 (2-59)，得到

$$\begin{aligned} r'(x)y(x)&+r(x)y'(x)\\ &=g(x)r(x)y(x)+[\theta(x)]^p\phi(x)y(x). \end{aligned}$$

由于 (2-57)，则简化成

$$r'(x)=[\theta(x)]^p\phi(x),$$

并且我们立得

$$r(x)=\int_a^x\theta(t)^p\phi(t)dt. \tag{2-61}$$

因此用简单的求积而无需知道真解便可能计算出相对伸缩误差．

作为一个例子，我们考虑方程 $y'=-16xy$ 的积分，在初值条件[1]

$$y(-0.75)=2^{-\frac{3}{2}}\pi^{-1/2}e^{-9/2}=0.0022159242$$

下，对定步长 h 使用 Runge-Kutta 方法，其精确解为

$$y(x)=2^{-\frac{3}{2}}\pi^{-\frac{1}{2}}e^{-8x^2}.$$

为了紧凑起见，令 $c=-16$，有

$$\begin{aligned} f(x,y)&=cxy,\\ f'(x,y)&=(c+c^2x^2)y,\\ f''(x,y)&=(3c^2x+c^3x^3)y, \end{aligned}$$

---

1) 这样选取的目的是可以使用 $(2\pi)^{-1/2}e^{-x^2/2}$ 的函数表(国家标准局[1942]).

$$f'''(x,y) = (3c^2 + 6c^3x^2 + c^4x^4)y,$$
$$f^{(\mathrm{IV})}(x,y) = (15c^3x + 10c^4x^3 + c^5x^5)y.$$

由 (2-36)，我们求得

$$\begin{aligned}\varphi(x,y) = & \frac{1}{2880}(15c^3x + 10c^4x^3 + c^5x^5)y \\ & - \frac{1}{576}cx(3c^2 + 6c^3x^2 + c^4x^4)y \\ & + \frac{1}{288}(c^2x^2 - c)(3c^2x + c^3x^3)y \\ & + \frac{1}{192}(2c^2x - 2c^3x^3)(c + c^4x^2)y = -\frac{c^5x^5}{120}y.\end{aligned}$$

于是给出的主误差函数正如所期望的那样，它也是 y 的线性函数. 从 (2-61) 容易得到

$$r(x) = -\frac{c^5}{720}(x^6 - a^6),$$

其中 $a = -0.75$. 它是 x 的偶函数，相对误差在以原点为对称点上近似相等. 由于精确解也是对称的，故对于绝对误差有相同结论成立. 因为 $r(\pm a) = 0$，所得 $y(-a)$ 具有四阶以上精确度.

在表 2.3 中，我们给出用步长 $h = 2^{-p}$，对于 $p = 4(1)9$，

表 2.3

x	-0.50	-0.25	0	0.25	0.50	0.75
$p = 4$	0.02693857	0.12069994	0.19899916	0.12069933	0.02694366	0.00222524
5	0.02699083	0.12096227	0.19943300	0.12096226	0.02699099	0.00221620
6	0.02699515	0.12098372	0.19946843	0.12098372	0.02699516	0.00221593
7	0.02699546	0.12098525	0.19947096	0.12098525	0.02699546	0.00221592
8	0.02699548	0.12098535	0.19947113	0.12098535	0.02699548	0.00221592
9	0.02699548	0.12098536	0.19947114	0.12098536	0.02699548	0.00221592

表 2.4

	$x=0$		$x=0.25$		$x=0.50$	
	真 正	预 估	真 正	预 估	真 正	预 估
$p=4$	−0.00047198	−0.00079376	−0.00028604	−0.00047785	−0.00005183	−0.0000974
5	0.00003814	0.00004961	0.00002311	0.00002987	0.00000449	0.00000608
6	0.00000271	0.00000310	0.00000164	0.00000187	0.00000033	0.00000038
7	0.00000018	0.00000019	0.00000011	0.00000012	0.00000002	0.00000002
8	0.00000001	0.00000001	0.00000001	0.00000001	0.00000000	0.00000000
9	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000

表 2.5

x	−0.50	−0.25	0	0.25	0.50	0.75
Euler $h=2^{-11}$	−0.00013981	−0.00078297	−0.00116196	−0.00062636	−0.00017476	−0.00002582
改进 Euler $h=2^{-10}$	−0.00000085	−0.00000428	−0.00000669	−0.00000428	−0.00000085	−0.00000000

计算出对应于[1] $x = -0.5(.25)0.75$ 的近似值 y_n. 每一列中最后的值与精确解所给出的数字是一致的.

在表 2.4 中比较真正误差与由理论预估的近似误差

$$e_n = h^4 r(x_n) y(x_n).$$

在表 2.3 的最后一列中，较快的数值收敛性在某种意义上证明了在 $x = 0.75$ 处零误差的预见是正确的.

为了便于比较，我们在表 2.5 中给出用 Euler 方法 (1-1) 和改进 Euler 方法(2—10b)得到的对同一个初值问题解的近似值的误差. 对于 Euler 方法，我们采用步长 $h = 2^{-11}$，对改进 Euler 方法，则用步长 $h = 2^{-10}$. 这些步长要求像用步长 $h = 2^{-9}$ 的 Runge-Kutta 方法一样多的函数 $f(x, y)$ 的求值次数，对于全部所考虑的 x 的值，有 $|e_n| < 10^{-8}$.

2.3. 一般单步方法的舍入误差

2.3-1. 一个先验界. 我们仍用 1.4-1 中所做的假设，并且继续使用那里引用的记号. 如果采取定点运算，代替(2-1)中的数值近似 $\tilde{y}_n$ 满足关系式

$$\tilde{y}_{n+1} = \tilde{y}_n + (h\tilde{\Phi}(x_n, \tilde{y}_n; h))^*, \qquad (2\text{-}62)$$

其中 $\tilde{\Phi}$ 表示 Φ 的舍入后的值，也可能为 Φ 的其它近似值. 我们把 (2-62) 写成形式

$$\tilde{y}_{n+1} = \tilde{y}_n + h\Phi(x_n, \tilde{y}_n; h) + \varepsilon_{n+1}, \qquad (2\text{-}63)$$

其中

$$\varepsilon_{n+1} = (h\tilde{\Phi}(x_n, \tilde{y}_n; h))^* - h\Phi(x_n, \tilde{y}_n; h) \qquad (2\text{-}64)$$

是局部舍入误差. 如前，把 ε_{n+1} 看成是固有误差

$$\rho_{n+1} = h[\tilde{\Phi}(x_n, \tilde{y}_n; h) - \Phi(x_n, \tilde{y}_n; h)] \qquad (2\text{-}65)$$

---

1) 如果 $c - a$ 是 b 的整数倍，记号 $a(b)c$ 表示数 $x = a + nb$ 的集合，其中 $n = 0, 1, \cdots, (c - a)b^{-1}$.

与引入误差

$$\pi_{n+1}=(h\bar{\Phi}(x_n,\tilde{y}_n;h))^*-h\bar{\Phi}(x_n,\tilde{y}_n;h) \quad (2-66)$$

的和.

本节将建立累积舍入误差 $r_n=\tilde{y}_n-y_n$ 的界，在单独假设

$$|\varepsilon_n|\leqslant\varepsilon,\quad n=1,2,\cdots \quad (2-67)$$

下,其中 ε 为常量. 从 (2-63) 减去理论上近似的对应的递推关系式 (2-1)，求得

$$r_{n+1}=r_n+h[\Phi(x_n,\tilde{y}_n;h)-\Phi(x_n,y_n;h)]+\varepsilon_{n+1}. \quad (2-68)$$

如果 L 表示方法的 Lipschitz 常数，由于 (2-15) 及 (2-67)，我们得到递推不等式

$$|r_{n+1}|\leqslant(1+hL)|r_n|+\varepsilon.$$

利用引理 1.1，解出这个递推关系式,我们得到:

定理 2.8. 如果增量函数 $\Phi(x, y; h)$ 满足条件 (2-15)，并且局部舍入误差以 (2-67) 为界,那么累积舍入误差满足

$$|r_n|\leqslant\frac{\varepsilon}{h}E_L(x_n-a),\quad a\leqslant x_n\leqslant b. \quad (2-69)$$

对于各种特殊方法的常数 L 值已在 2.2-3中给出.

上述结果的意义在于估计式 (2-69) 不依赖于常数 N 和 p，而它对于方法的离散误差具有代表性. 这就说明舍入误差的传播与方法的离散误差没有关系.

2.3-2. **累积舍入误差对局部舍入误差的依赖性.** 虽然估计式(2-69)在理论上有价值,但却没有多大的实际意义,因为它是基于这样的非现实假设之上，即所有局部误差取它们的最大值并且总是系统地互相加大. 还有，每一步都采用了 Lipschitz 估计 (2-15)，这就大大超过所提及的两个量之间的真实偏差.

作为对舍入误差较为实际估计的一个基础，我们研究它对局部舍入误差的依赖关系． 我们的目的是从所给关系式(2-68)出发，导出类似于(1-64)的一个关系式。仍然假设Φ有$p+1$阶的连续有界导数，我们有(因为$p\geqslant 1$)

$$\begin{aligned}&\Phi(x_n,\tilde{y}_n;h)-\Phi(x_n,y_n;h)\\&=\Phi_y(x_n,y(x_n);h)r_n+\Phi_{yy}(x_n,y^+;h)\frac{(\tilde{y}_n-y(x_n))^2}{2}\\&\quad-\Phi_{yy}(x_n,y^{++};h)\frac{(y_n-y(x_n))^2}{2},\end{aligned}\tag{2-70}$$

其中y^+与y^{++}是某些中间值。把Φ_y按h的幂次展开且利用

$$\Phi(x,y;0)=f(x,y).$$

令$g(x)=f_y(x,y(x))$，我们求得

$$\Phi_y(x_n,y(x_n);h)=g(x_n)+h\Phi_{yh}(x_n,y(x_n);h^+),$$

其中$0<h^+<h$。

为了得到有效的结果，必须令局部舍入误差的界ε按如下方式依赖于h:

$$\varepsilon\leqslant Kh^{q+1},\quad h<h_0,\tag{2-71}$$

其中q和K均与h和$q\geqslant 1$无关。我们还假设

$$Nh^{p+1}\leqslant\varepsilon,\tag{2-72}$$

其中N是出现在局部离散误差界(2-19)中的常数． 对于一切充分小的h值，仅当$q\leqslant p$时，方程(2-71)与(2-72)是相容的．

当局部舍入误差支配局部离散误差时，以上假设都是合适的．如果这个方法精确到使得局部离散误差小于基本单位u，便会自动地发生这种情形．

根据定理2.2，2.3和2.8，以上假设保证

$$|\tilde{y}_n-y_n|-|r_n|\leqslant\frac{\varepsilon}{h}E-h^qKE,$$

$$|\tilde{y}_n - y(x_n)| = |r_n + e_n| \leqslant \frac{2\varepsilon}{h} E = 2h^q KE, \quad x_n \in [a, b]$$

成立，其中 $E = E_L(b-a)$. 用M表示Φ的二阶导数的上界，于是由 (2-70) 推出

$$\Phi(x_n, \tilde{y}_n; h) - \Phi(x_n, y_n; h) = g(x_n) r_n + \theta_{n+1} \varepsilon K_1,$$

其中 $|\theta_n| \leqslant 1, n = 0, 1, \cdots$, 并且

$$K_1 = ME\left(1 + \frac{5}{2} h_0^{q-1} KE\right).$$

于是 (2-68) 可换成关系式

$$r_{n+1} = r_n + hg(x_n) r_n + \theta_{n+1} h \varepsilon K_1 + \varepsilon_{n+1}.$$

这与 (1-63) 是相同的，只不过待定量 ε_{n+1} 取为待定量 $\theta_{n+1} h \varepsilon K_1 + \varepsilon_{n+1}$. 象 1.4-3一样，导出

$$r_n = r_n^{(1)} + r_n^{(2)}, \tag{2-73}$$

其中

$$r_n^{(1)} = \sum_{m=1}^{n} d_{n,m} \varepsilon_m, \tag{2-73a}$$

$$r_n^{(2)} = h \varepsilon K_1 \sum_{m=1}^{n} d_{n,m} \theta_m. \tag{2-73b}$$

系数 $d_{n,m}$ 是由关系式 (1-65) 确定 [取 $g(x) = f_y(x, y(x))$]. 正如 1.4-3 中所指出的，它们满足

$$d_{n,m} = d_m(x_n) + O(h),$$

其中

$$d_m(x) = e^{G(x) - G(x_m)}, \quad G(x) = \int_a^x g(t) dt.$$

r_n 的分量 $r_n^{(1)}$ 称为累积舍入误差主要成分. 它是用 Euler 方法对线性方程 $y' = g(x)$ 的数值积分中出现的误差，如果局部误差相同. 如果函数 $m(x)$ 由 (1-72) 确定，便有

$$h \sum_{m=1}^{n} d_{m,n} = m(x_n) + O(h), \quad x_n \in [a, b].$$

对于充分小的 h 值，$d_{n,m}>0$. 如果 (2-67) 成立，这就导出

$$|r_n^{(1)}|\leqslant\frac{\varepsilon}{h}\{m(x_n)+O(h)\}. \tag{2-74}$$

分量 $r_n^{(2)}$ 称为累积舍入误差的次要成分. 我们把它的出现归结为所给微分方程的非线性以及用 Φ 确定方法的复杂性. 但是，利用上面推理，导出

$$|r_n^{(2)}|\leqslant\varepsilon K_1\{m(x_n)+O(h)\}. \tag{2-75}$$

(2-74) 和 (2-75) 的结果表明，在局部舍入误差支配局部离散误差的假设下，当 $h\to0$ 时，累积的主要舍入误差支配累积的次要舍入误差. 对所有单步方法，主要舍入误差的传播是相同的.

2.3-3. 可变局部舍入误差的一个后验界. 对于十进制 (或二进制) 定点运算，假设(2-67)是适当的，这里参与计算的所有数都是取相同的精确度. 为了放宽这些严格的假设，我们用以下更一般的条件来代替 (2-67)，

$$|\varepsilon_n|\leqslant\varepsilon p(x_n). \tag{2-76}$$

这里 $p(x)$ 为已知非负的 x 分段光滑函数. 上面所提到的性质是指区间 $[a,b]$ 可划分为有限个子区间，函数 $p(x)$ 在每一个子区间内是连续的且有连续导数，并且在两个端点处取有限极限. 量 ε 仍服从条件 (2-71) 及 (2-72). 在 2.3-5 中安排了条件 (2-76) 的一个应用；在 3.4-8 中讨论了浮点运算的另一个应用.

如果要对 $r_n^{(1)}$ 作估计，我们必须估计和

$$m_n=h\sum_{m=1}^{n}d_{n,m}p_m,$$

其中 $p_m=p(x_m)$. 我们通过建立一个差分方程来求关于 m_n 的近似表达式. 如果 $p(x)$在 $x=x_n$ 处可微，从而利用(1-65)，

有

$$
\begin{aligned}
m_{n+1} - m_n &= h\left\{\sum_{m=1}^{n+1} d_{n+1,m}p_m - \sum_{m=1}^{n} d_{n,m}p_m\right\} \\
&= h\left\{d_{n+1,n+1}p_{n+1} + \sum_{m=1}^{n} (d_{n+1,m} - d_{n,m})p_m\right\} \\
&= h\left\{p(x_{n+1}) + g(x_n)h\sum_{m=1}^{n} d_{n,m}p_m\right\} \\
&= h\{p(x_n) + g(x_n)m_n\} + O(h^2).
\end{aligned}
$$

最后关系式可以看成试图用 Euler 方法来解微分方程

$$m'(x) = g(x)m(x) + p(x). \qquad (2\text{-}77a)$$

根据假设，区间 $[a,b]$ 可划分成有限个子区间，$p(x)$ 在每一个子区间内连续．利用定理 1.4，从而在每个子区间上 m_n 近似于 (2-77a) 的一个适当的解．由此导出

$$m_n = m(x_n) + O(h), \qquad (2\text{-}78)$$

其中 $m(x)$ 是 (2-77a) 的连续解，且满足

$$m(a) = 0. \qquad (2\text{-}77b)$$

注意到 $r^{(2)} = O(\varepsilon)$，我们得到

定理 2.9. 如果局部舍入误差是以 (2-76) 为界，其中 $p(x)$ 是分段光滑函数且 ε 满足条件 (2-71) 和 (2-72)，那么累积舍入误差满足

$$|r_n| \leqslant \frac{\varepsilon}{h}\{m(x_n) + O(h)\}, \quad x_n \in [a,b], \qquad (2\text{-}79)$$

其中 $m(x)$ 由 (2-77) 所确定.

2.3-4. **统计估计.** 由于在 1.6 中所提供的数值结果，我们一定会料到除去很简单积分外，实际上界 (2-79) 大大地超过真正的舍入误差．用 1.5 和 1.6 中引入的统计方法，便可求数值积分中真正所期望的关于舍入误差的大小的真实结论．我们仍引用将局部舍入误差看成随机变量的假设．为了试验

性地验证这个假设，象以前一样，我们通过改变初值条件来得到对同一个微分方程的许多解.

在采用这个假设时，无需假设所有局部舍入误差是随机变量. 如果局部误差有一个成分不是随机变量，这个成分便可分出去并用前几节的方法来单独地估计. 但是，为了简化记号，我们仍用 ε_m 表示局部误差的随机部分.

象在第一章一样，我们假设变量 ε_m 有一个已知的分布. 利用 1.4 的结果，从而我们确定直接依赖于 ε_m 的累积舍入误差的 $r_n^{(1)}$ 部分的近似分布（关于 $r_n^{(2)}$ 一般不可能有统计结果，因为没有理由把 θ_n 当作随机量）. 但是，考虑到一般性，我们将允许舍入误差的分布可逐步地改变. 明确地说，我们假设

$$\begin{aligned} |E(\varepsilon_m)| &\leqslant \mu p(x_m), \\ \operatorname{var}(\varepsilon_m) &= \sigma^2 q(x_m), \end{aligned} \tag{2-80}$$

其中 $p(x)$ 和 $q(x)$ 为 $[a,b]$ 内已知非负分段光滑函数，非负量 μ 和 σ 都假设与 x 或 m 无关，虽然由于 (2-67) 和 (2-71) 而依赖于 h.

如果

$$r_n^{(1)} = \sum_{m=1}^{n} d_{n,m}\varepsilon_m,$$

利用 (1-94)，有

$$|E(r_n^{(1)})| \leqslant \mu \sum_{m=1}^{n} d_{n,m}p_m = \frac{\mu}{h} m_n,$$

或者由于 (2-78)，

$$|E(r_n^{(1)})| \leqslant \frac{\mu}{h}\{m(x_n) + O(h)\}.$$

如果 ε_m 都是独立的，利用 (1-95)，我们求得累积误差的方差

$$\operatorname{var}(r_n^{(1)}) = \sigma^2 \sum_{m=1}^{n} d_{n,m}^2 q_m.$$

从而变成去寻求和

$$v_n = h \sum_{m=1}^{n} d_{n,m}^2 q_m$$

的一个近似值． 由于(1-67)，量 v_n 对于 $x_n \in [a, b]$ 是有界的．

如果 $q(x)$ 在点 $x = x_n$ 处连续，再利用 $d_{n,m}$ 满足的差分方程(1-65)，我们求得

$$\begin{aligned} v_{n+1} - v_n &= h\left\{\sum_{m=1}^{n+1} d_{n+1,m}^2 q_m - \sum_{m=1}^{n} d_{n,m}^2 q_m\right\} \\ &= h\left\{q_{n+1} + \sum_{m=1}^{n} (d_{n+1,m}^2 - d_{n,m}^2) q_m\right\} \\ &= h\left\{q_{n+1} + \sum_{m=1}^{n} (d_{n+1,m} - d_{n,m})(d_{n+1,m} + d_{n,m}) q_m\right\} \\ &= h\left\{q_{n+1} + hg(x_n) \sum_{m=1}^{n} d_{n,m}^2 (2 + hg(x_n)) q_m\right\}. \end{aligned}$$

利用 $q(x)$ 和 v_n 的有界性，于是

$$v_{n+1} - v_n = h\{2g(x_n)v_n + q_{n+1}\} + O(h^2).$$

根据定理 1.4 推出，在 $q(x)$ 连续的每一个区间内，

$$v_n = v(x_n) + O(h), \tag{2-81}$$

其中 $V(x)$ 是微分方程

$$v'(x) = 2g(x)v(x) + q(x) \tag{2-82a}$$

的一个解．把在 $q(x)$ 不连续点上的解凑合成连续的，从而推出(2-81)在整个区间 $[a, b]$ 内成立，如果 $v(x)$ 是表示(2-82a)的连续且分段可微的解，并且满足

$$v(a) = 0. \tag{2-82b}$$

我们综合本节结果为以下定理．

定理 2.10． 令局部舍入误差 ε_n 为满足定理 2.9 条件的

随机变量且令它们的期望值和方差满足(2-80)，其中 $p(x)$ 和 $q(x)$ 为分段光滑函数，那么累积舍入误差的主要成分 $r_n^{(1)}$ 是一个随机变量，且有

$$|E(r_n^{(1)})| \leqslant \frac{\mu}{h}\{m(x_n)+O(h)\}, \tag{2-83}$$

$$\operatorname{var}(r_n^{(1)}) = \frac{\sigma^2}{h}\{v(x_n)+O(h)\}, x_n\in[a,b], \tag{2-84}$$

其中 $m(x)$ 和 $v(x)$ 分别由(2-77)和(2-82)所确定.

这个结果进一步证明了早先的结论，即单步方法的舍入误差不以任何方式与方法的截断误差有联系.

2.3-5. 部分双倍位精确度. 减少舍入影响的一个方法是完成所有的计算采用双倍位精确度. 从理论观点来看，这完全相当于以基本单位 u^2 来代替 u. 使舍入极小化的另一种措施，是使用所谓部分双倍位精确度. 这是一个简便的计算方法，用很少代价靠略微增加一些固有误差便可总体减少引入误差(把它看成局部舍入误差的主要来源). 这是通过用双倍位精确度计算 y_n 来达到的，而所有其它量用单倍位精确度来计算. 和前面一样，假设步长 h 和横坐标 x_n 都是精确的单倍位数.

象前面一样，用 z^* 表示(单倍位精确度)数 z 舍入后的值，部分双倍位精确度的算法由

$$\begin{aligned}&\tilde{y}_0 = y_0,\\&\tilde{y}_{n+1} = \tilde{y}_n + h\Phi(x_n, \tilde{y}_n^*; h),\quad n = 0,1,2,\cdots\end{aligned} \tag{2-85}$$

给出，这里 $\tilde{y}_n$ 是双倍位精确度的数. 这个算法的二个重要特点是:

(i) 乘积 $h\Phi$ 不带舍入且把它全部加到 $\tilde{y}_n$ 上;

(ii) 取 $\tilde{y}_n$ 更有效的部分来计算函数 Φ 值. 因此对计算 Φ 需要的时间并不比通常单倍位精确度的运算增多.

局部舍入误差 ε_{n+1} 规定为

$$\tilde{y}_{n+1} = \tilde{y}_n + h\Phi(x_n, \tilde{y}_n; h) + \varepsilon_{n+1},$$

于是在目前的情形，我们有

$$\varepsilon_{n+1} = h\tilde{\Phi}(x_n, \tilde{y}^*; h) - h\Phi(x_n, \tilde{y}_n; h).$$

为了估计 ε_{n+1}，我们写

$$\varepsilon_{n+1} = h[\tilde{\Phi}(x_n; \tilde{y}_n^*; h) - \Phi(x_n, \tilde{y}_n^*; h)] + h[\Phi(x_n, \tilde{y}_n^*; h) - \Phi(x_n, \tilde{y}_n; h)]. \quad (2\text{-}86)$$

如果所计算的函数 Φ 具有误差不超过一个固定量 $\delta>0$，那么有

$$|\varepsilon_{n+1}| \leqslant h\delta + \frac{1}{2} hLu,$$

或者更精确地，有

$$|\varepsilon_{n+1}| \leqslant h\delta + \frac{1}{2} hu[g(x_n) + O(h)]. \quad (2\text{-}87)$$

如果略去项 $O(h)$，最后的关系式便具有 (2-76) 形式，其中

$$\varepsilon = h\delta, \ \ p(x) = 1 + \frac{1}{2} u\delta^{-1}g(x).$$

关系式 (2-86) 还可作为统计处理的基础． 如果累积误差充分小，我们就有

$$\Phi(x_n, \tilde{y}_n^*; h) - \Phi(x_n, \tilde{y}_n; h) = g(x_n)(\tilde{y}_n^* - \tilde{y}_n) + O(h),$$

其中 $g(x) = f_y(x, y(x))$． 这就有理由假设量 $\tilde{y}_n^* - y_n$ 表现为具有由函数 $F_u(x)$ 确定的矩形分布（见 1.6-1)． 如果 $\tilde{\Phi} - \Phi$ 也作为随机变量，并假设它为

$$|E(\tilde{\Phi} - \Phi)| \leqslant mu, \ \ \mathrm{var}(\tilde{\Phi} - \Phi) = s^2u^2,$$

其中 m 和 s^2 都是常数[1]，那么推出 ε_{n+1} 是一个随机变量，使得到

$$|E(\varepsilon_{n+1})| \leqslant hmu, \quad (2\text{-}88)$$

---

1) 如果 $\tilde{\Phi} = \Phi^*$，我们可设 $m = 0$，$s^2 = \frac{1}{12}$.

并且如果变量 $\tilde{\Phi}-\Phi$ 和 $\tilde{y}^*-\tilde{y}$ 都是独立的，那么有

$$\operatorname{var}(\varepsilon_{n+1})=h^2\left(s^2+\frac{1}{12}g^2(x_n)\right)u^2+O(h^4u^2). \quad (2\text{-}89)$$

由此可见，对于部分双倍位精确度，即使采用二进制的定点运算，局部舍入误差的方差是依赖于 x 的．我们注意到，局部误差的方差的阶为 h^2u^2，这就说明了比通常的单倍位精确度改进了 h^2 倍． 从而累积舍入误差的标准偏差减小了与 h 同价的一个因子．

2.3-6．变步长．剩下来的是讨论对以下情形中的舍入误差的传播，即由公式

$$x_{n+1}=x_n+\theta(x_n)h \quad (2\text{-}90)$$

计算出一系列节点，其中 $\theta(x)$是与 h 或 n 无关的正的分段连续函数．如前，我们假设数 x_n 都是可以精确计算出来的．这个要求实际上限制 $\theta(x)$ 为分段常数函数类．但是，实际上当步长改变时便常常遇到这样情形．

从精确关系式

$$y_{n+1}=y_n+\theta(x_n)h\Phi(x_n,y_n;\theta(x_n)h)$$

减去数值所满足的对应关系式

$$\tilde{y}_{n+1}=\tilde{y}_n+\theta(x_n)h\Phi(x_n,\tilde{y}_n;\theta(x_n)h)+\varepsilon_{n+1}, \quad (2\text{-}91)$$

其中 ε_{n+1} 表示局部舍入误差，我们得到累积舍入误差的递推关系式．正如 2.3-2 中一样进行线性化，并且对 ε_{n+1} 用同样的假设，我们求得

$$r_{n+1}=r_n+h\theta(x_n)g(x_n)r_n+\eta_{n+1}, \quad (2\text{-}92)$$

其中

$$\eta_{n+1}=\varepsilon_{n+1}+\theta_{n+1}h\varepsilon K_1.$$

进一步的分析类似于(但不完全相同) 1.4-3． 对于任意的 η_m 值，假设

$$r_n=\sum_{m=1}^{n}d_{n,m}\eta_m, \quad (2\text{-}93)$$

把它代入到(2-92)，我们发现系数 $d_{n,m}$ 必须满足

$$d_{n+1,n+1}=1, n=0,1,2,\cdots, \tag{2-94}$$

$$d_{n+1,m}=d_{n,m}+h\theta(x_n)g(x_n)d_{n,m},$$

$$m=1,2,\cdots,\ n=m,m+1,\cdots.$$

利用定理 2.6 推出

$$d_{n,m}=d_m(x_n)+O(h) \tag{2-95}$$

其中

$$d'_m(x)=g(x)d_m(x),\ d_m(x_m)=1. \tag{2-96}$$

为了后验界和统计工作，我们必须计算和

$$m_n=h\sum_{m=1}^{n}d_{n,m}p_m, \tag{2-97}$$

其中 $p_m=p(x_m)$，$p(x)$的规定如 2.3-4。利用关系式(2-94)，容易求得公式

$$m_{n+1}=m_n+h[\theta(x_n)g(x_n)m_n+p(x_n)]+O(h^2),$$

又可以把它写成

$$m_{n+1}=m_n+\theta(x_n)h[g(x_n)m_n+\theta(x_n)^{-1}p(x_n)]+O(h^2).$$

利用定理 2.6，推出

$$m_n=m(x_n)+O(h), \tag{2-98}$$

其中

$$m'(x)=g(x)m(x)+\theta^{-1}(x)p(x),\ m(a)=0. \tag{2-99}$$

为了近似地确定累积误差的方差，需要计算

$$v_n=h\sum_{m=1}^{n}d_{n,m}^2q_m, \tag{2-100}$$

其中 $q_m=q(x_m)$. 如 2.3-4 中那样处理，我们求得

$$v_{n+1}=v_n+h[2\theta(x_n)g(x_n)v_n+q(x_n)]+O(h^2).$$

这个关系式可以写成形式

$$v_{n+1}=v_n+\theta(x_n)h[2g(x_n)v_n+\theta(x_n)^{-1}g(x_n)]+O(h^2),$$

并应用定理 2.6，推出

$$v_n = v(x_n) + O(h), \tag{2-101}$$

其中

$$v'(x) = 2g(x)v(x) + \theta^{-1}(x)q(x), \quad v(a) = 0. \tag{2-102}$$

一旦计算出和(2-97)与(2-100)，便如同2.3-3和2.3-4一样进行讨论. 从而我们得到如下的最终定理，它概括了本节中这种类型的综合结果.

定理 2.11. 令点(x_n, y_n)是由(2-52)确定的，并令局部舍入误差满足(2-76)，其中ε使得(2-71)和(2-72)成立，那么累积舍入误差满足

$$|r_n| \leqslant \frac{\varepsilon}{h}\{m(x_n) + O(h)\}, \quad x_n \in [a, b], \tag{2-103}$$

其中$m(x)$由(2-99)确定. 此外，如果局部舍入误差是满足(2-80)的互相独立的随机变量，那么累积舍入误差的主要成份$r_n^{(1)}$是一个随机变量，并且

$$|E(r_n^{(1)})| \leqslant \frac{\mu}{h}\{m(x_n) + O(h)\}, \tag{2-104}$$

$$\operatorname{var}(r_n^{(1)}) = \frac{\sigma^2}{h}[v(x_n) + O(h)], \quad x_n \in [a, b], \tag{2-105}$$

其中$m(x)$同上，$v(x)$由(2-102)确定.

在$m(x)$和$v(x)$的定义中出现项$\theta(x)^{-1}$，它暗示在步长小从而$\theta(x)^{-1}$大的区域对累积舍入误差的影响要比使用大的步长的区域来得大.

2.3-7. 数值例子. 前节的结果实质上是依赖于把局部舍入误差处理为随机变量的假设，并且这些结果涉及到方差，这就必须假设这些变量都是独立的. 我们将给出一些数值上的证据来支持这些假设，而不是试图去证明这些假设.

选取初值问题：

$$y' = -16xy,$$

$$y(-0.75) = y_{0,q} = y_{0,0}(1 + q\Delta), \quad q = 0, 1, \cdots, Q - 1,$$

的数值积分作为例子，其中

$$y_{0,0} = 0.0022159242 \cdot 2^{-14},$$

$$\Delta = \frac{1}{3} \cdot 2^{-8},$$

$$Q = 500.$$

采用 (I) 改进 Euler 方法；(II) Runge-Kutta 方法.

对应于初值为 $y_{0,q}$ 在点 x_n 处的数值和理论近似值分别以 $\tilde{y}_{n,q}$ 和 $y_{n,q}$ 来表示.

我们首先假设 h 是固定的，$h = 2^{-6}$，并且考虑在不同的 x 值上的舍入误差. 用二进制定点运算 ($u = 2^{-36}$)，很自然地假设局部舍入误差具有由函数 $F_u(x)$ 确定的固定的矩形分布(见 1.6-1). 从而推出

$$E(\varepsilon_m) = 0, \quad \mathrm{var}(\varepsilon_m) = \frac{1}{12} u^2.$$

因此对所考察的二个方法，希望有

$$E(r_m) = 0, \quad \mathrm{var}(r_n) = \frac{2^6}{12} u^2 V(x_n),$$

其中

$$v'(x) = -32xv(x) + 1, \quad v(-0.75) = 0.$$

在表 2.6 中给出 $v(x)$ 的值(由数值积分求得).

在表 2.7 (也可见图 2.2) 中对这些预估值与由公式

$$E(r_n) = \frac{1}{Q} \sum_{q=0}^{Q-1} (\tilde{y}_{n,q} - y_{n,q}),$$

$$\mathrm{var}(r_n) = \frac{1}{Q} \sum_{q=0}^{Q-1} (\tilde{y}_{n,q} - E(r_n))^2$$

数值地计算出的均值和方差作了比较. "精确" 值 $y_{n,q}$ 也可由数值计算得到，但要用较高的精确度来计算. 值得注意的

表 2.6

x	-0.50	-0.25	0	0.25	0.50	0.75
$v(x)$	6.5	132.7	361.1	133.0	6.7	0.1

表 2.7

x_n		-0.50	-0.25	0	0.25	0.50	0.75
$u^{-1}E(r_n)$	预估	0.00	0.00	0.00	0.00	0.00	0.00
	方法 I	0.71	3.37	5.83	3.48	0.68	0.04
	方法 II	1.05	5.02	7.78	5.02	1.03	0.05
$u^{-2}\mathrm{var}(r_n)$	预估	34.6	707.7	1925.8	709.3	35.7	0.5
	方法 I	42.6	844.1	2304.3	848.8	43.1	0.5
	方法 II	41.6	842.2	2282.6	841.6	41.6	0.5

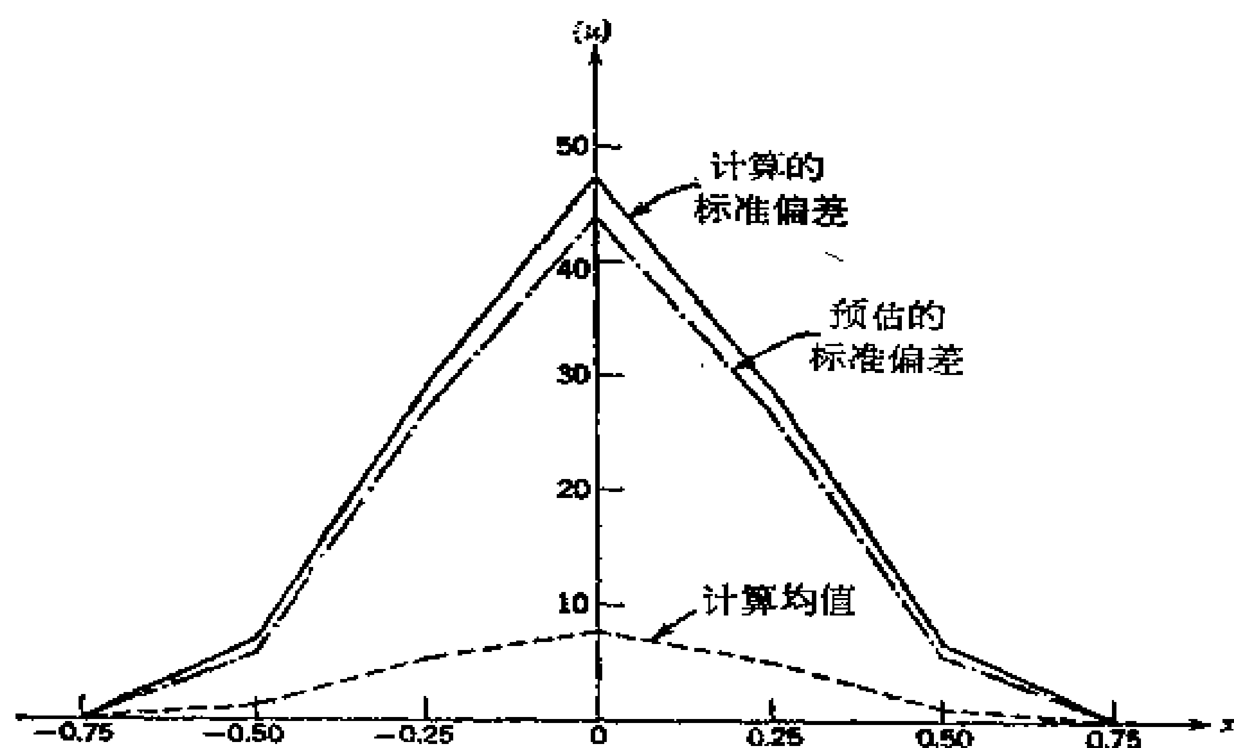


图 2.2 Runge-Kutta 方法的舍入误差传播

是，在 $x=0$ 处数值解有相当大的偏差，随后又在 $x=0.75$ 处会聚，这正和理论上所预估的一样.

保持 h 固定而让 x 变化，也可以研究在一个固定的值 x_n

上把数值 $y_{n,q}$ 的偏差作为 h 的一个函数. 在 $x_n=0$ 处预估方差为

$$u^{-2}\mathrm{var}(r_n)\approx\frac{1}{h}\cdot\frac{1}{12}\cdot 361.1.$$

在表 2.8 中试验值是用改进 Euler 方法来计算的. 对于大的 h 值,结果并不很好,但随着 h 减小却变得较好. 这可以用以下事实来阐明,即在上面的分析中,局部舍入误差假设只包含引入误差. 当 $h\to 0$ 时,这是渐近地成立的;但是,对于大的 h 值,固有误差却不能完全忽略.

表 2.8

h		2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}
$u^{-2}\mathrm{var}(r_n)$	预估	481.4	962.9	1925.9	3851.7	7703.5
	试验	848.9	1380.7	2304.6	4556.0	7801.0

2.4. 求解的问题

2.1

1. 解初值问题

$$y'=\frac{1}{1-0.2\cos y},\quad y(0)=0.$$

在区间 $[0,2\pi]$ 上:

(a) 采用改进 Euler 方法,取 $h=\pi/5$;

(b) 采用四阶 Runge-Kutta 方法,取 $h=2\pi/5$.

[(a) $y_{10}=6.27246$, (b) $y_5=6.28176$.]

2. 证明使用改进 Euler 方法可以精确地求解微分方程 $y'=-2ax$.

3. 关系式

$$y_{n+1}-y_n=\frac{1}{2}h[f(x_n,y_n)+f(x_{n+1},y_{n+1})] \qquad (2\text{-}106)$$

确定一个单步方法，虽因右端出现 y_{n+1}，但此时所确定的增量函数仅为隐式. 假设由 (2-106) 解出 y_{n+1}，并且这个解是形如 $y_{n+1} = y_n + c_1 h + c_2 h^2 + \cdots$ 的展式. 试确定这个方法的阶.

4. 证明由

$$y_{n+1} = y_n + \frac{1}{6} h[4f(x_n, y_n) + 2f(x_{n+1}, y_{n+1}) + hf'(x_n, y_n)]$$

确定的隐式单步方法的阶为 3.

5. 令 $g(x,y) = f'(x,y) = f_x + f_y f$. 证明由增量函数

$$\Phi(x,y;h) = f(x,y) + \frac{1}{2} hg\left(x + \frac{1}{3} h, y + \frac{1}{3} hf(x,y)\right)$$

确定的单步方法的阶为 3.

6. 考察由

$$\Phi(x,y;h) = f\left(x + \frac{1}{2} h, y + \frac{1}{2} h\bar{\Phi}\left(x,y;\frac{1}{2} h\right)\right) \tag{2-107}$$

确定的单步方法,其中 $\bar{\Phi}$ 表示另一个增量函数. 对于二次连续可微函数 $z(x)$，利用

$$h^{-1}\int_x^{x+h} z(t)dt = z\left(x + \frac{1}{2} h\right) + \frac{1}{24} h^2 z''(\xi),$$

其中 $x < \xi < x + h$，证明: 如果由 (2-107) 确定的方法的阶为正,那么由 (2-107) 确定的方法的阶至少为 2.

7. 如果 $z(t)$ 三次连续可微,那么

$$h^{-1}\int_x^{x+h} z(t)dt = \frac{1}{4} z(x) + \frac{3}{4} z\left(x + \frac{2}{3} h\right) + \frac{1}{216} h^3 z'''(\xi),$$

其中 $x < \xi < x + h$. 如果由增量函数 $\bar{\Phi}$ 确定的方法的阶至少为 2，证明由

$$\Phi(x,y;h)=\frac{1}{4}f(x,y)+\frac{3}{4}f\Big(x+\frac{2}{3}h,$$

$$y+\frac{2}{3}h\bar{\Phi}\Big(x,y;\frac{2}{3}h\Big)\Big)\qquad(2\text{-}108)$$

确定的方法的阶至少为 3.

8*. 建立在求积基础上的单步方法. 令常数 $w_k,\theta_k\in[0,1](k=1,2,\cdots,\nu)$, $C\neq0$,并且整数 μ 使得对任意充分可微函数 $z(t)$, 可选取一个数 $\xi\in(x,x+h)$, 使得以下求积公式成立:

$$h^{-1}\int_x^{x+h}z(t)dt=\sum_{k=1}^{\nu}w_kz(x+\theta_kh)+Ch^{\mu}z^{(\mu)}(\xi).$$

(2-109)

令函数 $\bar{\Phi}_k(x,y;h)(k=1,2,\cdots,\nu)$ 是任意增量函数, 证明由

$$\Phi(x,y;h)=\sum_{k=1}^{\nu}w_kf(x+\theta_kh,y+\theta_kh\bar{\Phi}_k(x,y;\theta_kh))$$

(2-110)

确定的单步方法的阶至少为 $\min(\mu,\bar{p}+1)$, 其中 $\bar{p}$ 表示由函数 $\bar{\Phi}$ 确定的方法的最小阶.

9*. 如果一般 Runge-Kutta 方法如显式单步方法一样来确定,在此方法中增量函数完全由函数 $f(x,y)$ 确定. 利用对于任意 μ 存在形式为(2-109)公式的事实, 证明存在任意阶的 Runge-Kutta 方法. [提示: 假设存在 n 阶的 Runge-Kutta 方法, 将它与 $\mu=n+1$ 的求积公式联在一起便产生一个 $n+1$ 阶方法.]

10*. 利用对于一切偶数 μ, 取 $\nu=\frac{1}{2}\mu$ 以及对于一切奇数 μ, 取 $\nu=\frac{1}{2}(\mu+1)$, $\theta_1=0$ 都存在一个求积公式的事

实 (Hildebrand [1956, Chap. VIII]). 证明对于 $2n$ 阶的 Runge-Kutta 方法所要求计算的 f 的次数不超过 $(n!)^2$.

2.2

11. 采用 Heun 方法 (2-10a) 确定初值问题

$$y' = -2xy, \quad y(-2) = e^{-4}$$

的数值解的伸缩误差函数,误差小于 $O(h^2)$ 的有哪些点?

12. 确定方法 (2-106) 的主误差函数, 并用精确解的导数来表示.

13. 用 $\varphi(x,y)$ 表示由 $\Phi(x,y;h)$ 确定的方法的主误差函数,导出由 (2-107) 给出的方法的主误差函数公式[区分 Φ 的阶 $\bar{p}$ 满足 $\bar{p}=1$ 和 $\bar{p}>1$ 的情形].

14*. 证明: 若对于问题 8 的方法 $\bar{p} \geqslant \mu$, 则主误差函数为

$$\varphi(x,y) = -Cf^{(\mu)}(x,y).$$

15. 证明: 对于问题 $y' = ay$, $y(0) = 1$ 改进 Euler 方法恒同于二阶 Taylor 方法,经典的 Runge-Kutta 方法恒同于 4 阶 Taylor 方法,并求主误差函数的近似表达式.

16. 对不同的 h 值, 采用同一个方法对固定区间上微分方程的积分的试验中, 假设离散误差(近似地)作为 h 的函数有如下的值:

$h = 2^{-4}$	$e_n = 533 \cdot 10^{-8}$
2^{-5}	$65 \cdot 10^{-8}$
2^{-6}	$8 \cdot 10^{-8}$
2^{-7}	$1 \cdot 10^{-8}$

你能说出关于这个方法的阶是多少吗?

17. 用步长 $h = 2^{-p} \cdot \dfrac{1}{9}$ $(p = 0,1,2,\cdots)$ 通常的 Euler 方法对问题 $y' = -3y$, $y(0) = \dfrac{1}{8}$ 的数值积分在 $x_* = 8/9$

处得到如下的值：

ν	y_ν
0	0.004877305
1	0.006760986
2	0.007721044
3	0.008202987
4	0.008444179
5	0.008564801
6	0.008625115
7	0.008655273
8	0.008670352
9	0.008677891
10	0.00867659

$\left(\text{精确结果：}\ y(8/9)=\frac{1}{8}e^{-\frac{8}{3}}\approx 0.00868543\right)$检验用外推到极限 $h=0$ 得到的值有出现为 $O(h^3)$ 而不是 $O(h^2)$ 的误差，这正如原先所期望的一样． 或者用展开 y_ν 为 h 的幂次或者用第一章的问题 8 的结果来说明之．

18. Heun [1900] 提出增量函数为

$$\Phi(x,y;h)=\frac{1}{4}(k_1+3k_3)$$

的 Runge-Kutta 方法，其中

$$k_{\nu+1}=f\left(x+\frac{1}{3}\nu h,\ y+\frac{1}{3}\nu h k_\nu\right),\quad \nu=0,1,2.$$

证明这是一个三阶方法，并且确定其主误差函数．

19. 从二种途径用 Runge-Kutta 方法来解初值问题

$$y'=-y\operatorname{sign}x,\quad y(-0.5)=0.1.$$

令

$$\Phi(x,y;h)=\sum_{\nu=1}^{4}a_\nu f(X_\nu,Y_\nu),$$

其中 a_ν, X_ν, Y_ν 由 (2-12b) 确定,或者令

$$f(X_\nu, Y_\nu) = \begin{cases} +Y_\nu, & \text{如果 } X_\nu < 0, \\ -Y_\nu, & \text{如果 } X_\nu \geqslant 0, \end{cases} \tag{i}$$

或者令

$$f(X_\nu, Y_\nu) = \begin{cases} +Y_\nu, & \text{如果 } x < 0, \\ -Y_\nu, & \text{如果 } x \geqslant 0. \end{cases} \tag{ii}$$

对于步长 $h = 2^{-p}$, $p = 2, 3, \cdots$, 在点 $x_n = 0.5$ 得到以下数值:

p	y_n	
	(i)	(ii)
2	0.091659898	0.10000683
3	0.095838792	0.10000021
4	0.097916628	0.10000000
5	0.09895838330	0.10000000
6	0.0994791699	
7	0.099739583	
8	0.099934895	
9	0.099967447	
10	0.099983723	

验证在情形 (i), 误差阶为 h 且说明之. [在情形 (ii) 中, 对于这个特殊情形的误差阶为 h^5, 虽然通常只希望为 h^4 阶.]

20*. 对给定步数的误差极小化. 如果用 $\theta(x)$ 修改步长所确定的变步长,并用主误差函数为 $\varphi(x,y)$ 的 k 阶单步方法来解初值问题 (1-1), 那么解出 (2-56) 便容易看出伸缩误差为

$$e(x) = u(x)\int_a^x u(t)^{-1}\varphi(t, y(t))[\theta(t)]^k dt, \tag{2-111}$$

其中 $$u(x) = \exp\int_a^x f_y(t, y(t))dt.$$

如果只要求对定值 $x=b$ 的解，把（2-111）当作 $\theta(t)$ 的函数，考虑（2-111）的极小化问题，在边界条件

$$\int_a^b \frac{1}{\theta(t)} dt = 1$$

下．对于小的 h 值，它近似地表达步数为常数的事实［函数 $\theta(t)^{-1}$ 是节点的"密度"］．假设 $\varphi(t, y(t)) > 0$，证明这个极小值问题的解[1]为

$$\theta(t) = \frac{A}{g(t)},$$

其中

$$g(t) = [u(t)^{-1}\varphi(t, y(t))]^{\frac{1}{k+1}},$$

$$A = \int_a^b g(t) dt.$$

作为一个应用，证明：对任意 b，问题 $y' = y$, $y(0) = 1$ 取均匀步长为最好解．

［提示：利用 Hölder 不等式

$$\int_a^b \alpha\beta\, dt \leqslant \left(\int_a^b \alpha^p dt\right)^{1/p} \left(\int_a^b \beta^q dt\right)^{1/q}.］$$

2.3

21．在关于局部舍入误差常用的统计假设下，使用步长 $h = 2^{-8}$ 的单精确度的运算，用单步方法在 2.3-7 讨论数值问题中的舍入误差超过：

(i) $175u$ 在 $x=0$，(ii) $2.8u$ 在 $x = 0.75$ 的概率（近似地）是多少？

22．浮点运算的舍入．如果 $y' = g(x)y$ 以及 $p = y$, $q = y^2$，证明分别由（2-77）和（2-82）所确定的函数 $m(x)$ 和

1）这个问题的解是 D. D. Morrison 给出的。

$v(x)$也可定义为二阶方程

$$m'' = 2gm' + (g' - g^2)m$$

及
$$v'' = 4gv' + (2g' - 4g^2)v$$

的解，且满足初值条件

$$m(a) = 0, \quad m'(a) = y(a)$$

及
$$v(a) = 0, \ v'(a) = [y(a)]^2.$$

还要证明在相同的条件下，函数

$$\mu(x) = \frac{m(x)}{y(x)}, \quad w(x) = \frac{v(x)}{y^2(x)}$$

（舍入误差的相对量度）满足

$$\mu(x) = w(x) = x - a.$$

作为一个应用，比较用浮点和定点运算的单步法对 $y' = -y$，$y(0) = 1$ 的解的舍入误差传播.

23. Runge-Kutta 方法的固有舍入误差。令经典的 Runge-Kutta 的增量函数写成形式

$$\Phi = \frac{1}{6}K_1 + \frac{1}{3}K_2 + \frac{1}{3}K_3 + \frac{1}{6}K_4,$$

Φ 的数值或由

$$\Phi = \left[\left(\frac{1}{6}\right)^* (K_1^* + 2K_2^* + 2K_3^* + K_4^*)\right]^* \tag{A}$$

或由

$$\Phi = \left[\left(\frac{1}{6}\right)^* K_1^*\right]^* + \left[\left(\frac{1}{3}\right)^* K_2^*\right]^* + \left[\left(\frac{1}{3}\right)^* K_3^*\right]^* + \left[\left(\frac{1}{6}\right)^* K_4^*\right]^* \tag{B}$$

计算出来.

假设量 $K_\nu^* - K_\nu$ 与舍入误差其乘积取无穷多位二进制小数都是随机变量，其均值为 0 而方差为 $\sigma^2 = \frac{1}{12}u^2$，证明在

情形（A），

$$E(\bar{\Phi}-\Phi)=\left[\left(\frac{1}{6}\right)^*-\frac{1}{6}\right]\Phi,\quad \operatorname{var}(\bar{\Phi}-\Phi)=\frac{23}{216}u^2;$$

在情形（B），

$$E(\bar{\Phi}-\Phi)=0,\quad \operatorname{var}(\bar{\Phi}-\Phi)=\frac{77}{216}u^2,$$

如果 $\left(\frac{1}{6}\right)^*+\left(\frac{1}{3}\right)^*+\left(\frac{1}{3}\right)^*+\left(\frac{1}{6}\right)^*=1$.

（这个结果对部分双倍位精确度是重要的.）

注

2.1-2．关于 Runge-Kutta 方法的经典参考文献为 Runge [1895]，Heun [1900]，Kutta [1901]．各种变形方法是 Gill [1951]，Albrecht [1955]，Conte 和 Reeves [1956]，Blum [1957]，Collatz [1960，p. 64] 提供的． 把 Runge-Kutta 方法的思想与使用高阶导数相结合这是 Zurmühl [1948] 和 Fehlberg [1958] 提出的． 超过四阶的 Runge-Kutta 方法是 Nyström [1925] 和 Huta [1956，1957] 给出的． 在数字计算机上的 Runge-Kutta 方法的程序设计是 Murray [1950]，Gill [1951]，Blum [1957]，Martin [1958]，Romanelli [1960] 研究的．在许多文章中涉及到建立在积分公式基础上的单步方法．Milne [1949] 使用含有高阶导数的求积公式；以及 Lotkin [1952]（见问题 3 和 4）采用十分相似的方法，也使用了这样公式．所有这些公式是 Obrechkoff [1942] 的基本公式的应用结果．对常微分方程数值积分所使用的 Gauss 或类似求积公式的思想出现在 Hammer 和 Hollingsworth [1955]，Morrison 和 Stoller [1958]，Korganoff [1958]（见问题 6，7，8，13，14）．文献的大部分是由 U. S. S. R. 出版．在同一个时期，Caplygin 给出一个方法的误差界，见 Babkin [1948]，Luzin [1951]，

Azbelev [1952, 1953, 1955], Voronovskaya [1955]. Weissinger [1953] 有一个隐式微分方程的方法. 关于其它方法, 见 Vlasow 和 Čarnyi [1950], Bukovics [1950], Bückner [1952].

2.2-2, 2.3-3, 2.2-4. 对于经典的 Runge-Kutta 方法, 单个方程的局部离散误差的界是 Bieberbach [1944 p. 54] 和 Lotkin [1951] 给出的. 误差传播是 Bukovics [1953, 1954], Ceschino [1954], Carr [1958] 研究的. Milne [1950] 给出一个数值例子.

2.2-6, 2.2-7, 2.2-8. 用于单步方法及多步方法的离散误差的渐近性态启示性处理见 Lotkin [1954]. 估计局部舍入误差的方法是 Gorn 和 Moore [1953], Morel [1956], Kuntzmann [1959a] 给出的. Garfinkel [1954], Ceschino [1956] 处理了变步长.

2.3-5. 部分双倍位精确度的方法是 Young [1955] 提供的. "控制"舍入误差的其它方法由 Gill [1951] 和 Blum [1957] 给出.

第三章　一阶方程组的一般单步方法

3.1. 理论上的介绍

3.1-1. 定义. 一阶常微分方程组是形如

$$
\left.\begin{aligned}
y^{1\prime} &= f^1(x, y^1, y^2, \cdots, y^s)\\
y^{2\prime} &= f^2(x, y^1, y^2, \cdots, y^s)\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
y^{s\prime} &= f^s(x, y^1, y^2, \cdots, y^s)
\end{aligned}\right\} \tag{3-1}
$$

的方程组，其中函数 $f^1, f^2, \cdots, f^s$ 都是它们的 $s+1$ 个变元的已知函数(在本章中，上标总是表示指标而不是乘方). 函数集合 $y^1(x), y^2(x), \cdots, y^s(x)$ 称为方程组的解，如果它们在区间 $[a,b]$ 上是确定的并且是可微的而且对 x 恒满足关系式

$$
y^{i\prime} = f^i(x, y^1(x), y^2(x), \cdots, y^s(x)),\quad i = 1, 2, \cdots, s. \tag{3-2}
$$

实际中常常发生且在本章中将要讨论的问题是：求满足给定初值条件为

$$
y^i(a) = \eta^i,\quad i = 1, 2, \cdots, s \tag{3-3}
$$

的解，其中数 η^i 是预先指定的常数. 这个问题称为初值问题. 比 (3-3) 更为复杂的其它条件在实际中也会发生.

常微分方程组产生于以下几个方面.

(i) 在理论上，高于一阶的每一个常微分方程均可化成一阶方程组. 给定 m 阶微分方程

$$
y^{(m)} = f(x, y, y', y'', \cdots, y^{(m-1)}), \tag{3-4}
$$

其中 f 是它的 $m+1$ 个变元的已知函数(象通常一样，括号里

的上标表示导数)．令

$$y = y^1, y' = y^2, \cdots, y^{(m-1)} = y^m, \qquad (3\text{-}5)$$

便可将它化成方程组．如果函数 $y^1, y^2, \cdots, y^m$ 满足方程组

$$\begin{aligned} y^{1'} &= y^2, \\ y^{2'} &= y^3, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ y^{m'} &= f(x, y^1, y^2, \cdots, y^m), \end{aligned} \qquad (3\text{-}6)$$

这是(3-1)的特殊情形，那么函数 $y(x) = y^1(x)$ 显然满足(3-4)．作为一个例子，我们考虑二阶方程

$$y'' = -y.$$

令

$$y = y^1, \quad y' = y^2,$$

便化成方程组，于是方程组形如

$$\begin{aligned} y^{1'} &= y^2, \\ y^{2'} &= -y^1. \end{aligned} \qquad (3\text{-}7)$$

为了数值上的目的，有些作者(Milne [1953], p.82; Gill [1951], p.96)建议把高阶方程化为一阶方程组；另一些作者(Collatz [1960], p.117)却反对这样做，而认为化成一阶方程组既增长误差又增加必要的运算次数．在第四、六章中提供并研究高阶方程的直接积分方法．但是，在这些章中所提供的理论和实验的结果表明，当把高阶方程首先化成方程组，然后用一个适当方法解这个方程组时，其截断误差一般来说是不增长的，并且舍入误差实质上却常常是减少的．

(ii) 在许多物理问题中也自然地产生常微分方程组．典型的例子是多于一个环路的电路以及多个自由度的力学问题．更为明显的例子是陀螺仪的运动方程，外弹道基本方程组以及控制火箭飞行的方程．

(iii) 近来，还使用常微分方程组来得到偏微分方程的近

似解（Anonymous [1957], p.73). 为了说明这个方法，我们应用它来求满足偏微分方程

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2},\ 0\leqslant x<L,\ t\geqslant 0 \tag{3-8}$$

以及边界条件

$$u(x,0)=f(x),\ 0\leqslant x\leqslant L$$
$$u(0,t)=g(t),\ u(L,t)=h(t),\ t\geqslant 0$$

的函数 $u(x,t)$ 的问题，其中 $f(t)$，$g(t)$ 和 $h(t)$ 都是已知函数.

对于这个问题的通常数值逼近就是既在 x 方向又在 t 方向把微分方程 (3-8) 离散化. 这里所讨论的方法与此不同，它仅对这二个变量中的一个变量进行离散化，而对另一个变量却保留问题的微分特征. 如果我们选取在 x 方向离散，那么对于每一个 $t\geqslant 0$ 及 $x=x_n=nk$ $(k=L/N, n=1,2,\cdots,N-1)$，以

$$\frac{1}{k^2}\left(u(x_{n+1},t)-2u(x_n,t)+u(x_{n-1},t)\right)$$

近似 $\partial^2 u/\partial x^2$. 令

$$u^n(t)=u(x_n,t)$$

且保留在 t 方向的微分特征，从而我们可用常微分方程组

$$\frac{du^n}{dt}=k^{-2}(u^{n+1}-2u^n+u^{n-1}),\ n=1,2,\cdots,N-1$$

来代替 (3-8)，其中 $u^0=g(t)$，$u^N=h(t)$. 初值条件是

$$u^n(0)=f(x_n),\ n=1,2,\cdots,N-1.$$

这个方法容易应用于更为复杂的偏微分方程的问题. 关于误差估计，见 Rothe [1930], Budak [1956], Douglas [1956], Franklin [1959].

3.1-2. 向量记号. 把量 $y^i(i=1,2,\cdots,s)$ 看成为向量

$$\mathbf{y}=\begin{pmatrix}y^1\\y^2\\\vdots\\y^s\end{pmatrix}$$

的分量，这在概念及形式上都大为简化了以后的分析．因此，我们还令 $f^i(x, y^1, y^2, \cdots, y^s)=f^i(x, \mathbf{y})$． 把这 s 个函数 $f^i(x,y)$ 组成另一个向量

$$\mathbf{f}(x,\mathbf{y})=\begin{pmatrix}f^1(x, \mathbf{y})\\f^2(x, \mathbf{y})\\\vdots\\f^s(x, \mathbf{y})\end{pmatrix},$$

我们便可把组（3-1）写成更为紧凑的形式

$$\mathbf{y}'=\mathbf{f}(x,\mathbf{y}). \qquad (3\text{-}9\text{a})$$

如果规定向量 $\boldsymbol{\eta}$ 为

$$\boldsymbol{\eta}=\begin{pmatrix}\eta^1\\\eta^2\\\vdots\\\eta^s\end{pmatrix},$$

则初值条件（3-3）是

$$\mathbf{y}(a)=\boldsymbol{\eta}. \qquad (3\text{-}9\text{b})$$

我们需要类似于实或复数的绝对值的一个向量．如果 $\mathbf{v}$ 是一个向量，它具有实或复的分量 $v^i(i=1, 2, \cdots, s)$，在本章中采用如下的定义：

$$\|\mathbf{v}\|=|v^1|+|v^2|+\cdots+|v^s|. \qquad (3\text{-}10)$$

我们注意到，对于任意实的或复的数 λ 及任意两个向量 $\mathbf{v}$ 和 $\mathbf{w}$，有以下关系式成立：

$$\|\mathbf{v}\|\geqslant 0;\ \|\mathbf{v}\|=0,\ \text{当且仅当}\ \mathbf{v}=\mathbf{0} \qquad (3\text{-}11\text{a})$$

$$\|\lambda\mathbf{v}\| = |\lambda|\|\mathbf{v}\|, \tag{3-11b}$$

$$\|\mathbf{v}+\mathbf{w}\| \leqslant \|\mathbf{v}\| + \|\mathbf{w}\| \text{（三角不等式）}, \tag{3-11c}$$

这里向量与纯量的乘积以及二个向量的和都是按通常方法来定义的.

(3-11)描绘出数 $\|\mathbf{v}\|$ 为向量 $\mathbf{v}$ 的范数. 还可用范数的其它定义(在第七章就是这样做的),但是对于当前的问题,上述的定义则更为方便.

如果 $\mathbf{v}_n(n=0,1,2,\cdots)$ 表示一个向量序列,那么记号

$$\lim_{n\to\infty}\mathbf{v}_n = \mathbf{w} \text{ 意即 } \lim_{n\to\infty}\|\mathbf{v}_n - \mathbf{w}\| = 0,$$

或者等价于序列 v_n^i 有极限 $w^i(i=1,2,\cdots,s)$.

如果向量 $\mathbf{v}$ 的每一个分量 v^i 都依赖于变量 t,严格地说,便称 $\mathbf{v}$ 为 t 的向量值函数. 但是,如果不会发生误解,常常略去"向量值"这个词. 而

$$\lim_{t\to t_0}\mathbf{v}(t) = \mathbf{z} \text{ 是意指 } \lim_{t\to t_0}\|\mathbf{v}(t) - \mathbf{z}\| = 0.$$

如果

$$\lim_{t\to t_0}\mathbf{v}(t) = \mathbf{v}(t_0),$$

便称函数 $\mathbf{v}(t)$ 在 $t=t_0$ 处是连续的. 用 $\mathbf{v}'(t)$ 与 $\int_a^b \mathbf{v}(t)dt$ 表示这样的向量,其分量分别为 $\mathbf{v}(t)$ 的分量的导数与积分. 如果对于 $t\to t_0$, $\varphi(t)\to 0$, 那么

$$\mathbf{v}(t) = O(\varphi(t)),\quad t\to t_0,$$

意即

$$\|\mathbf{v}(t)\| = O(\varphi(t)),\quad t\to t_0.$$

对于依赖于多个变量的向量 $\mathbf{v}$,其极限、连续及阶的概念均可作类似的定义.

3.1-3. *初值问题解的存在性.* 我们假设纯变量 x 和向量 $\mathbf{y}=(y^1, y^2,\cdots,y^s)$ 的向量值函数 $\mathbf{f}(x,\mathbf{y})$ 满足以下两个

假设:

(A) $\mathbf{f}(x,\mathbf{y})$ 在区域 $a \leqslant x \leqslant b$, $-\infty < y^i < +\infty$, $i = 1,2,\cdots,s$ 内是确定的且连续的;

(B) 存在一个 Lipschitz 常数 L, 使得: 对于任意

$$x \in [a,b]$$

和任意二个向量 $\mathbf{y}$ 同 $\mathbf{y}^*$, 有

$$\|\mathbf{f}(x,\mathbf{y}) - \mathbf{f}(x,\mathbf{y}^*)\| \leqslant L\|\mathbf{y} - \mathbf{y}^*\|.$$

我们将证明类似于定理 1.1 的如下定理:

定理 3.1. 令函数 $\mathbf{f}(x, \mathbf{y})$ 满足上述条件 (A) 和 (B), 并令 $\boldsymbol{\eta}$ 为已知向量, 那么恰好存在具有如下三个性质的一个函数 $\mathbf{y}(x)$:

(i) $\mathbf{y}(x)$对 $x \in [a,b]$ 是连续的且连续可微;

(ii) $\mathbf{y}'(x) = \mathbf{f}(x,\mathbf{y})$, $x \in [a,b]$;

(iii) $\mathbf{y}(a) = \boldsymbol{\eta}$.

简言之,初值问题 (3-9) 有唯一解.

定理 3.1 的证明与定理 1.1 的证明十分相似. 我们从构造一个函数序列 $\mathbf{y}_p(x)(p = 0,1,2,\cdots)$ 开始,证明它收敛于一个连续函数 $\mathbf{y}(x)$. 这个构造与 §§1.2-2 至 1.2-5 中给出的构造极其相似,故仅作概述. 然后我们证明 $\mathbf{y}(x)$是初值问题 (3-9) 的解. 可以采用 §§1.2-6 和 1.2-7 中十分初等的方法来证明这个结论. 但是由于各种原因, 我们在这里将用不同的且不甚初等的方法来证明.

3.1-4. 近似解 $\mathbf{y}_p(x)$ 的收敛性. 如果 h_p 和 $x_{[p]}$ 如 1.2-2 中所定义, 对于 $p=0, 1, 2, \cdots$, 我们规定函数 $\mathbf{y}_p(x)$ 为

$$\mathbf{y}_p(a) = \boldsymbol{\eta}, \tag{3-12}$$

$$\mathbf{y}_p(x) = \mathbf{y}_p(x_{[p]}) + (x - x_{[p]})\mathbf{f}(x_{[p]},\mathbf{y}_p(x_{[p]})).$$

这些函数在每一个区间 $(x_{[p]}, x_{[p]} + h_p)$ 内显然都是连续的,

并且根据其构造，在点 $x_{[p]}$ 处也是连续的. 我们将证明类似于引理 1.1 的如下引理:

引理 3.1. 向量函数序列 $\mathbf{y}_p(x)$ 当 $p\to\infty$ 时对 $x\in[a,b]$ 一致收敛于一个连续函数 $\mathbf{y}(x)$.

证. 利用条件 (B), 取 $\mathbf{y}^*=0$, 我们求得

$$\|\mathbf{f}(x,\mathbf{y})\|\leqslant L\|\mathbf{y}\|+c, \tag{3-13}$$

其中

$$c=\max_{x\in[a,b]}\|\mathbf{f}(x,0)\|.$$

从 (3-12), 得到

$$\mathbf{y}_p(x_{[p]}+h_p)=\mathbf{y}_p(x_{[p]})+h_p\mathbf{f}(x_{[p]},\mathbf{y}_p(x_{[p]})),$$
$$x\in(a,b-h_p].$$

利用 (3-13), 于是

$$\|\mathbf{y}_p(x_{[p]}+h_p)\|\leqslant(1+h_pL)\|\mathbf{y}_p(x_{[p]})\|+h_pc.$$

应用引理 1.2, 取 $\xi_n=\|\mathbf{y}(a+nh_p)\|$, $A=(1+h_pL)$, $B=h_pc$, 导出

$$\|\mathbf{y}_p(x_{[p]})\|\leqslant e^{(x_{[p]}-a)L}\|\boldsymbol{\eta}\|+E_L(x_{[p]}-a)c,$$

其中 E_L 表示 Lipschitz 函数. 由于函数 $\mathbf{y}_p(x)$ 在任意两个相邻点 $x_{[p]}$ 之间都是线性的,我们有

$$\|\mathbf{y}_p(x)\|\leqslant Y,\quad x\in[a,b], \tag{3-14}$$

其中

$$Y=e^{(b-a)L}\|\boldsymbol{\eta}\|+E_L(b-a)c. \tag{3-15}$$

我们用 R 表示由 $x\in[a,b]$, $\|\mathbf{y}\|\leqslant Y$ 定义的 $(x,\mathbf{y})$ 空间内的紧致区域,从而不等式 (3-14) 表明,对于 $x\in[a,b]$ 和

$$p=0,1,2,\cdots,$$

函数 $\mathbf{y}_p(x)$ 都在 R 内. 由于 R 的紧致性, 属于 R 内的连续函数在 R 内有有限的最大值. 我们令

$$M=\max_{(x,\mathbf{y})\in R}\|\mathbf{f}(x,\mathbf{y})\|, \tag{3-16}$$

此外，对于任意 $\delta \geqslant 0$，令

$$\omega(\delta) = \max\|\mathbf{f}(x,\mathbf{y}) - \mathbf{f}(x^*,\mathbf{y})\|, \qquad (3\text{-}17)$$

其中最大值是对 R 内使得 $|x - x^*| \leqslant \delta$ 的一切点 $(x, \mathbf{y})$ 和 $(x^*,\mathbf{y})$ 来取的. 正如在 §1.2-4 中一样，我们有

$$\lim_{\delta\to 0}\omega(\delta) = 0. \qquad (3\text{-}18)$$

为了应用 Cauchy 法则，我们令

$$\mathbf{d}(x) = \mathbf{y}_p(x) - \mathbf{y}_q(x),$$

其中 p 和 q 是任意两个非负整数，$p < q$. 类似于引理 1.3，下述引理必然是多维的.

引理 3.2. 对于 $t \in [a,b]$，则有

$$\|\mathbf{d}(t)\| \leqslant [1 + (t - t_{[q]})L]\|\mathbf{d}(t_{[q]})\| + (t - t_{[q]})\Omega_p, \qquad (3\text{-}19)$$

其中

$$\Omega_p = \omega(h_p) + LMh_p.$$

证. 象证明引理 1.3 一样，我们可写成

$$\begin{aligned}\mathbf{d}(t) - \mathbf{d}(t_{[q]}) = (t - t_{[q]})\{&\mathbf{f}(t_{[q]},\mathbf{y}_q(t_{[q]}))\\ &- \mathbf{f}(t_{[q]},\mathbf{y}_p(t_{[q]})) + \mathbf{f}(t_{[q]},\mathbf{y}_p(t_{[q]}))\\ &- \mathbf{f}(t_{[p]},\mathbf{y}_p(t_{[q]})) + \mathbf{f}(t_{[p]},\mathbf{y}_p(t_{[q]}))\\ &- \mathbf{f}(t_{[p]},\mathbf{y}_p(t_{[p]}))\}.\end{aligned}$$

右端前二行中的函数 $\mathbf{f}$ 值的差可分别用 $L\|\mathbf{d}(t_{[q]})\|$ 及 $\omega(h_{[p]})$ 来估计. 利用 (B)，对最后一个差的范数，我们有估计

$$L\|\mathbf{y}_p(t_{[q]}) - \mathbf{y}_p(t_{[p]})\|.$$

或者由于 (3-12)，并利用 $t_{[q]}-t_{[p]} \leqslant h_p$，这个估计为 LMh_p. 引理 3.2 的结论明显成立.

如同从 (1-24) 导出 (1-27)，也完全可推出不等式

$$\|\mathbf{d}(x)\| \leqslant \Omega_p\frac{e^{(x-a)L} - 1}{L}. \qquad (3\text{-}20)$$

由于 (3-20) 右端的表达式不依赖于 q，并且利用 (3-18)，当

$p \to \infty$ 时，它对 $x \in [a,b]$ 一致地趋向零，从而推出引理 3.1 的结果。

3.1-5. 完成存在性定理的证明. 为了证明极限函数 $\mathbf{y}(x)$ 满足所给定的微分方程，对于 $p=0,1,2,\cdots$，我们规定向量值函数 $\mathbf{f}_p(x)$（依赖于单个纯变量 x）为

$$\mathbf{f}_p(x)=\begin{cases}\mathbf{f}(a,\boldsymbol{\eta}), & x=a,\\ \mathbf{f}(x_{[p]},\mathbf{y}_p(x_{[p]})), & a<x\leqslant b.\end{cases} \tag{3-21}$$

函数 $\mathbf{f}_p(x)$ 在每一个区间 $(x_{[p]},x_{[p]}+h_p]$ 内都是常数，并且每一个分量的常数值是对应于该区间内 $\mathbf{y}_p(x)$ 的分量的斜率. 因此

$$\mathbf{y}_p(x)-\boldsymbol{\eta}=\int_a^x \mathbf{f}_p(t)\,dt. \tag{3-22}$$

我们将在下面证明

$$\lim_{p\to\infty}\mathbf{f}_p(t)=\mathbf{f}(t,\mathbf{y}(t)) \tag{3-23}$$

对于 $t\in[a,b]$ 一致成立. 假设这个事实成立，我们在 (3-22) 中可令 $p\to\infty$，得

$$\mathbf{y}(x)-\boldsymbol{\eta}=\int_a^x \mathbf{f}(t,\mathbf{y}(t))\,dt. \tag{3-24}$$

条件 (A) 指出 $\mathbf{f}(t,\mathbf{y}(t))$ 是连续的，因此在 (3-24) 中积分的导数存在且等于 $\mathbf{f}(x,\mathbf{y}(x))$. 从而左端函数的导数也存在且有相同的值. 因此

$$\mathbf{y}'(x)=\mathbf{f}(x,\mathbf{y}(x)),\quad x\in[a,b].$$

正如所望.

剩下来的是证明 (3-23). 我们有

$$\begin{aligned}\|\mathbf{f}_p(t)-\mathbf{f}(t,\mathbf{y}(t))\| &\leqslant \|\mathbf{f}(t_{[p]},\mathbf{y}_p(t_{[p]}))-\mathbf{f}(t,\mathbf{y}_p(t_{[p]}))\|\\ &\quad+\|\mathbf{f}(t,\mathbf{y}_p(t_{[p]}))-\mathbf{f}(t,\mathbf{y}_p(t))\|\\ &\quad+\|\mathbf{f}(t,\mathbf{y}_p(t))-\mathbf{f}(t,\mathbf{y}(t))\|\\ &\leqslant \omega(h_p)+LMh_p+L\|\mathbf{y}_p(t)-\mathbf{y}(t)\|.\end{aligned}$$

选取 p 足够大，最后表达式对 t 可一致地任意小。这就证明了(3-23)。

为了证明 $\mathbf{y}(x)$ 是初值问题的唯一解，我们注意到，按定义，初值问题的任意解在区间 $[a,b]$ 上是连续的，从而它是有界的。如果 $\mathbf{y}(x)$ 及 $\mathbf{z}(x)$ 是任意两个解，则它们的差

$$\boldsymbol{\delta}(x)=\mathbf{y}(x)-\mathbf{z}(x)$$

也是有界的：

$$\|\boldsymbol{\delta}(x)\|\leqslant K,\quad x\in[a,b].\tag{3-25}$$

由

$$\mathbf{y}(x)-\boldsymbol{\eta}=\int_a^x\mathbf{f}(t,\mathbf{y}(t))dt,$$

与

$$\mathbf{z}(x)-\boldsymbol{\eta}=\int_a^x\mathbf{f}(t,\mathbf{z}(t))dt$$

相减，导出

$$\boldsymbol{\delta}(x)=\int_a^x[\mathbf{f}(t,\mathbf{y}(t))-\mathbf{f}(t,\mathbf{z}(t))]dt.$$

利用 Lipschitz 条件，于是

$$\|\boldsymbol{\delta}(x)\|\leqslant L\int_a^x\|\boldsymbol{\delta}(t)\|dt.\tag{3-26}$$

现在我们证明，对于 $k=0,1,2,\cdots$，有

$$\|\boldsymbol{\delta}(x)\|\leqslant K\frac{L^k(x-a)^k}{k!},\quad x\in[a,b].\tag{3-27}$$

根据(3-25)，这个不等式对于 $k=0$ 是成立的。假设对于一个正整数 k 它是成立的，利用(3-26)，我们求得

$$\|\boldsymbol{\delta}(x)\|\leqslant L\int_a^x\frac{KL^k(t-a)^k}{k!}dt=K\frac{L^{k+1}(x-a)^{k+1}}{(k+1)!}.$$

这就证明了(3-27)中的 k 增加1。从而推出对一切非负整数 k，(3-27)是成立的。因为

$$\frac{L^k(b-a)^k}{k!}\to 0,\quad k\to\infty$$

仅当 $\|\boldsymbol{\delta}(x)\|=0$，$a\leqslant x\leqslant b$ 才有可能，所以导出

$$\mathbf{y}(x)\equiv\mathbf{z}(x).$$

这便完成了定理 3.1 的证明.

可应用存在性定理的一个重要特殊情形，即线性微分方程组

$$\mathbf{y}'=\mathbf{G}(x)\mathbf{y}+\mathbf{q}(x), \tag{3-28}$$

其中

$$\mathbf{G}(x)=\begin{pmatrix} g^{11}(x) & g^{12}(x) & \cdots & g^{1s}(x)\\ g^{21}(x) & g^{22}(x) & \cdots & g^{2s}(x)\\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots\\ g^{s1}(x) & g^{s2}(x) & \cdots & g^{ss}(x)\end{pmatrix}$$

是 $s\times s$ 矩阵，其元素都是 x 的连续函数，$x\in[a,b]$；并且 $\mathbf{q}(x)$ 也是连续的. 因此满足条件 (B)，取

$$L=\max_{x\in[a,b]}\|\mathbf{G}(x)\|,$$

其中

$$\|\mathbf{G}(x)\|=\max_{1\leqslant j\leqslant s}\sum_{i=1}^{s}|g^{ij}(x)|.$$

3.2. 方程组的特殊单步方法

3.2-1. 定义；记号简化. 令 $x\in[a,b]$，并令 $\mathbf{y}$ 是任意一个向量. 用 $\mathbf{z}(t)$ 表示满足 $\mathbf{z}(x)=\mathbf{y}$ 的微分方程组 $\mathbf{z}'=\mathbf{f}(t,\mathbf{z})$ 的解，我们令

$$\boldsymbol{\Delta}(x,\mathbf{y};h)=\begin{cases}\dfrac{\mathbf{z}(x+h)-\mathbf{z}(x)}{h}, & h\neq 0,\\ \mathbf{f}(x,\mathbf{y}), & h=0,\end{cases} \tag{3-29}$$

并且称 $\mathbf{\Delta}$ 为给定的微分方程解的精确相对增量. 由公式

$$\mathbf{y}_0 = \boldsymbol{\eta},$$
$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\boldsymbol{\Phi}(x,\mathbf{y}_n;h), \quad n = 0,1,\cdots \tag{3-30}$$

给出初值问题(3-9)的解的单步方法. 在这里选取函数 $\boldsymbol{\Phi}(x,\mathbf{y};h)$ 使之尽可能近似于 $\mathbf{\Delta}(x,\mathbf{y};h)$, 并且称 $\boldsymbol{\Phi}(x,\mathbf{y};h)$ 为增量函数. 如果 p 是使得

$$\boldsymbol{\Phi}(x,\mathbf{y};h) - \mathbf{\Delta}(x,\mathbf{y};h) = O(h^p) \tag{3-31}$$

成立的最大整数, 那么称 p 是由(3-30)所确定的方法的(精确)阶. 在叙述增量函数一些性质之前,我们将引入一个记号上的技巧,它可以简化以后的分析.

为了取消自变量 x 的特殊作用,我们用一个新函数 $y^0(x)$ 所满足的微分方程

$$y^{0\prime} = 1 \tag{3-32}$$

添加到组(3-1)中,并且满足初值条件

$$y^0(a) = a. \tag{3-33}$$

显然由(3-32)及(3-33)导出 $y^0(x) = x$, 因此组(3-1)可用 $s+1$ 个函数的等价组

$$y^{i\prime} = f^i(y^0,y^1,\cdots,y^s), \quad i = 0,\cdots,s \tag{3-34}$$

来代替,并且默认

$$f^0(y^0,y^1,\cdots,y^s) = 1. \tag{3-35}$$

这个新的组(3-34)具有这样的优点,即函数 f^i 中所含的变量都可看成是因变量.

为了简化,我们仍用 $y^1,\cdots,y^s$ 来表示因变量, 不管它们之中有 x 或没有 x. 于是我们把初值问题(3-9)写成如下形式:

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}), \ \mathbf{y}(a) = \boldsymbol{\eta}, \tag{3-36}$$

其中 $\mathbf{y}$, $\mathbf{f}$ 及 $\boldsymbol{\eta}$ 仍是具有 s 个分量的向量. 如果 $\mathbf{f}(\mathbf{y})$ 不明显地依赖于 x, 我们就可认为函数 $\mathbf{\Delta}$ 与增函数 $\boldsymbol{\Phi}$ 也不依赖于

x. 于是我们令

$$\boldsymbol{\Delta} = \boldsymbol{\Delta}(\mathbf{y};h), \quad \boldsymbol{\Phi} = \boldsymbol{\Phi}(\mathbf{y};h).$$

现在我们来构造一些特殊的增量函数 $\boldsymbol{\Phi}$.

3.2-2. *利用 Taylor 展式；张量记号.* 如果函数 $\mathbf{y}(x)$ 是 (3-36) 的解，并且 $\mathbf{f}$ 的分量都是充分可微的，那么 $\mathbf{y}(x)$ 的高阶导数可用函数 $\mathbf{f}$ 及其导数来表示. 例如

$$\mathbf{y}''(x) = \frac{d}{dx}\mathbf{f}(y(x)) = \sum_{j=1}^{s}\frac{\partial \mathbf{f}}{\partial y^j}\cdot\frac{dy^j}{dx} = \sum_{j=1}^{s}\frac{\partial \mathbf{f}}{\partial y^j}f^j.$$

一般地，由于可微性的假设，对于 $k = 1,2,\cdots$，我们有

$$\mathbf{y}^{(k+1)}(x) = \frac{d^k}{dx^k}\mathbf{f}(\mathbf{y}(x)) = \mathbf{f}^{(k)}(\mathbf{y}(x)).$$

函数 $\mathbf{f}^{(k)}$ 是由向量 $\mathbf{y}$ 完全确定的函数，从而我们有

$$\boldsymbol{\Delta}(\mathbf{y};h) = \mathbf{f}(\mathbf{y}) + \frac{h}{2}\mathbf{f}'(\mathbf{y}) + \frac{h^2}{3!}\mathbf{f}''(\mathbf{y}) + \cdots. \quad (3-37)$$

象单个方程的情形一样，近似积分方法可建立在截断 Taylor 级数的基础上. 对于 p 阶 Taylor 展式方法，其增量函数为

$$\boldsymbol{\Phi}(\mathbf{y};h) = \mathbf{f}(\mathbf{y}) + \frac{h}{2}\mathbf{f}'(\mathbf{y}) + \cdots + \frac{h^{p-1}}{p!}f^{(p-1)}(\mathbf{y}). \quad (3-38)$$

当 $p = 1$ 时，它便化成 Euler 方法.

由 (3-38) 定义的方法并没有很大的实际意义，因为计算导数 $\mathbf{f}^{(k)}$ 一般来说是很复杂的[1]. 但是，由于以下事实，这个方法在理论上却是重要的. 下面所讨论的 Runge-Kutta 型方法都是基于这样的思想，即通过不含有不同于 $\mathbf{f}(\mathbf{y})$ 的任何函数的表达式来近似 (3-38). 由于这个原因，我们将详细考察导数 $\mathbf{f}^{(k)}(\mathbf{y})$ 的结构.

---

1) 特殊微分方程的情形除外，见 Milley [1955], Cheney [1960].

为了简化记号，对于 $i,j,k=1,\cdots,s$，我们令

$$\frac{\partial f^i}{\partial y^j}=f^i_j,\quad \frac{\partial^2 f^i}{\partial y^j\partial y^k}=f^i_{jk},\tag{3-39}$$

并且对高阶导数也使用类似记号．所谓 $f_j, f_{jk},\cdots$，是指具有分量 $f^i_j, f^i_{jk},\cdots(i=1,\cdots,s)$ 的向量．我们还约定，如果一个标号既出现在上标又出现在下标，则这个乘积便是对于这个标号从 1 到 s 求和．例如

$$f^i_j f^j=\sum_{j=1}^{s}\frac{\partial f^i}{\partial y^j}f^j.\tag{3-40}$$

容易验证，形如(3-40)的乘积的和可以象通常乘积一样来求微分．因此

$$\begin{aligned}\frac{d}{dx}(f^i_j f^j)&=\left(\frac{d}{dx}f^i_j\right)f^j+f^i_j\left(\frac{d}{dx}f^j\right)\\&=f^i_{jk}f^j f^k+f^i_j f^j_k f^k.\end{aligned}$$

我们引入缩写记号：

$$\begin{aligned}&A^i=f^i,\quad E^i=f^i_{jkm}f^j f^k f^m,\\&B^i=f^i_j f^j,\quad F^i=f^i_{jk}f^j_m f^k f^m,\\&C^i=f^i_{jk}f^j f^k,\quad G^i=f^i_j f^j_{km}f^m f^k,\\&D^i=f^i_j f^j_k f^k,\quad H^i=f^i_j f^j_k f^k_m f^m,\end{aligned}\tag{3-41}$$

其中所有标号都是从 1 到 s 以及所有函数的变元都理解为 $\mathbf{y}$，并且用 $\mathbf{A}$、$\mathbf{B}$、$\cdots$表示分量为 A^i，B^i，$\cdots(i=1,2,\cdots,s)$ 的向量．

利用这些记号上的技巧，前几个导数 $\mathbf{f}^{(k)}$ 在形式上（仅是形式上）可用紧凑方式表达于下：

$$\begin{aligned}&\mathbf{f}(\mathbf{y})=\mathbf{A},\\&\mathbf{f}'(\mathbf{y})=\mathbf{B},\\&\mathbf{f}''(\mathbf{y})=\mathbf{C}+\mathbf{D},\\&\mathbf{f}'''(\mathbf{y})=\mathbf{E}+3\mathbf{F}+\mathbf{G}+\mathbf{H}.\end{aligned}\tag{3-42}$$

我们把 $\mathbf{f}(\mathbf{y}+h\mathbf{a})$ 展开成 h 幂次的表达式，其中 $\mathbf{y}$ 和

$$\mathbf{a}=\begin{pmatrix} a^1 \\ a^2 \\ \vdots \\ a^s \end{pmatrix}$$

都是固定向量．对多变元函数应用 Taylor 定理的结果便可写成形式

$$f^i(\mathbf{y}+h\mathbf{a})=f^i+hf^i_j a^j+\frac{1}{2}h^2 f^i_{jk}a^j a^k+\frac{1}{6}h^3 f^i_{jkl}a^j a^k a^l$$
$$+O(h^4). \tag{3-43}$$

3.2-3. Runge-Kutta 型方法． 我们讨论仅是间接基于 Taylor 展式的方法．象 §2-12 一样，这个思想就是把不同点上函数 $\mathbf{f}$ 的值(而非它的导数)组合起来．用这样的方法使所得到的函数 $\boldsymbol{\Phi}(\mathbf{y};h)$ 尽可能与 (3-37) 或者按照新记号与

$$\mathbf{A}+\frac{1}{2}h\mathbf{B}+\frac{1}{6}h^2(\mathbf{C}+\mathbf{D})$$
$$+\frac{1}{24}h^3(\mathbf{E}+3\mathbf{F}+\mathbf{G}+\mathbf{H})+\cdots \tag{3-44}$$

相一致． 我们通过讨论具有两次代换的“简化” Runge-Kutta 方法来阐明这个解析方法．我们试探性地令

$$\boldsymbol{\Phi}(\mathbf{y};h)=a_1\mathbf{f}(\mathbf{y})+a_2\mathbf{f}(\mathbf{y}+ph\mathbf{f}(\mathbf{y})),$$

其中常数 a_1, a_2 及 p 都是待定的．利用 (3-43)，我们有

$$\mathbf{f}(\mathbf{y}+ph\mathbf{f}(\mathbf{y}))=\mathbf{f}(\mathbf{y})+hp\mathbf{f}_j(\mathbf{y})f^j(\mathbf{y})$$
$$+\frac{1}{2}(hp)^2\mathbf{f}_{jk}(\mathbf{y})f^j(\mathbf{y})f^k(\mathbf{y})+O(h^3)$$
$$=\mathbf{A}+hp\mathbf{B}+\frac{1}{2}(hp)^2\mathbf{C}+O(h^3).$$

从而

$$\boldsymbol{\Phi}(\mathbf{y};h)=(a_1+a_2)\mathbf{A}+a_2hp\mathbf{B}+\frac{1}{2}a_2(hp)^2\mathbf{C}$$
$$+O(h^3). \tag{3-45}$$

使其常数项和 h 的线性项与(3-44)中对应的项相等，我们得到条件

$$a_1 + a_2 = 1,$$
$$a_2 p = \frac{1}{2}. \tag{3-46}$$

它不可能使二次项一致，因为在(3-45)中缺少向量 **D**. (3-46)的通解为

$$a_1 = 1 - \alpha, \quad a_2 = \alpha, \quad p = \frac{1}{2\alpha},$$

其中 $\alpha \neq 0$. 于是所需要的增量函数为

$$\boldsymbol{\Phi}(\mathbf{y}; h) = (1 - \alpha)\mathbf{f}(\mathbf{y}) + \alpha \mathbf{f}\left(\mathbf{y} + \frac{h}{2\alpha}\mathbf{f}(\mathbf{y})\right), \quad \alpha \neq 0. \tag{3-47}$$

它与(3-44)相差 $O(h^2)$. 为了计算 $\boldsymbol{\Phi}$，需要计算向量函数 $\mathbf{f}(\mathbf{y})$ 两次.

现在我们试图看一下计算四次函数将会怎样. 我们规定四个向量如下:

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{y}),$$
$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y} + hp_1\mathbf{k}_1),$$
$$\mathbf{k}_3 = \mathbf{f}(\mathbf{y} + h(p_2 - p_3)\mathbf{k}_1 + hp_3\mathbf{k}_2),$$
$$\mathbf{k}_4 = \mathbf{f}(y + h(p_4 - p_5 - p_6)\mathbf{k}_1 + hp_5\mathbf{k}_2 + hp_6\mathbf{k}_3),$$

并且试图确定常数 $p_1, \cdots, p_6$ 及 $a_1, \cdots, a_4$，使得表达式

$$\boldsymbol{\Phi}(\mathbf{y}; h) = a_1\mathbf{k}_1 + a_2\mathbf{k}_2 + a_3\mathbf{k}_3 + a_4\mathbf{k}_4 \tag{3-48}$$

与(3-44)尽可能地一致. 通过重复使用(3-43)，展开成 h 的幂次，并略去 $O(h^4)$，我们求得

$$\mathbf{k}_1 = \mathbf{A},$$

$$\mathbf{k}_2 = \mathbf{A} + hp_1\mathbf{B} + \frac{1}{2}h^2p_1^2\mathbf{C} + \frac{1}{6}h^3p_1^3\mathbf{E},$$

$$
\begin{aligned}
\mathbf{k}_3 = \mathbf{A} + hp_2\mathbf{B} + \frac{1}{2}h^2(p_2^2\mathbf{C} + 2p_1p_3\mathbf{D}) \\
+ \frac{1}{6}h^3(p_2^3\mathbf{E} + 6p_1p_2p_3\mathbf{F} + 3p_1^2p_3\mathbf{G}),
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{k}_4 = \mathbf{A} + hp_4\mathbf{B} + \frac{1}{2}h^2[p_4^2\mathbf{C} + 2(p_1p_5 + p_2p_6)\mathbf{D}] \\
+ \frac{1}{6}h^3[p_4^3\mathbf{E} + 6p_4(p_1p_5 + p_2p_6)\mathbf{F} \\
+ 3(p_1^2p_5 + p_2^2p_6)\mathbf{G} + 6p_1p_3p_6\mathbf{H}].
\end{aligned}
$$

代人到(3-48)中并使向量 $\mathbf{A},\cdots,\mathbf{H}$ 的系数与(3-44)中相应的系数相等,我们得到关于十个参数 $a_1,\cdots,a_4,p_1,\cdots,p_6$ 如下的 8 个方程:

$$
\begin{aligned}
a_1 + a_2 + a_3 + a_4 &= 1, \\
a_2p_1 + a_3p_2 + a_4p_4 &= \frac{1}{2}, \\
a_2p_1^2 + a_3p_2^2 + a_4p_4^2 &= \frac{1}{3}, \\
a_3p_1p_3 + a_4(p_1p_5 + p_2p_6) &= \frac{1}{6}, \\
a_2p_1^3 + a_3p_2^3 + a_4p_4^3 &= \frac{1}{4}, \\
a_3p_1p_2p_3 + a_4p_4(p_1p_5 + p_2p_6) &= \frac{1}{8}, \\
a_3p_1^2p_3 + a_4(p_1^2p_5 + p_2^2p_6) &= \frac{1}{12}, \\
a_4p_1p_3p_6 &= \frac{1}{24}.
\end{aligned}
\qquad (3\text{-}49)
$$

[对于 $a_3 = a_4 = 0$,前两个方程化成(3-46),这与前面相同.] 上面非线性方程组的通解是困难的. Kutta [1901] 给出如下一个参数解族:

a_1	a_2	a_3	a_4
$\frac{1}{6}$	$\frac{2-t}{3}$	$\frac{t}{3}$	$\frac{1}{6}$

(3-50)

p_1	p_2	p_3	p_4	p_5	p_6
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2t}$	1	$1-t$	t

直到最近，对应于 $t=1$ 的解还是最通用的。当然，这是经典 Runge-Kutta 情形，其中

$$a_1=a_4=\frac{1}{6},\quad a_2=a_3=\frac{1}{3}.$$

并且对于向量 $\mathbf{k}_j(j=1,2,3,4)$ 的公式有简单形式

$$\begin{aligned}\mathbf{k}_1&=\mathbf{f}(\mathbf{y}),\\ \mathbf{k}_2&=\mathbf{f}\left(\mathbf{y}+\frac{1}{2}h\mathbf{k}_1\right),\\ \mathbf{k}_3&=\mathbf{f}\left(y+\frac{1}{2}h\mathbf{k}_2\right),\\ \mathbf{k}_4&=\mathbf{f}(y+h\mathbf{k}_3).\end{aligned}\tag{3-51}$$

在 §2.1-2 中对单个微分方程的积分给出的 Runge-Kutta 公式是这个结果的特殊情形。

Gill [1951] 建议使用 $t=1-2^{-\frac{1}{2}}$ 的 Kutta 解 (3-50)。为了在那个时候限于机器贮存量而采用这种选取。Gill 指出，使用这个特殊选取的 t，为了实现 Runge-Kutta 方法同时贮存 s 个分量的四个向量变为线性相关，于是真正要贮存的仅为三个向量。另一方面，Gill 选取的 t 使之增加了复杂性 (p_5 不再为零) 并且系数失去对称性。此外，Blum [1957] 指出，如有必要，经典 Runge-Kutta 方法也可达到节省同样贮存量。

关于涉及到 Runge-Kutta 方法计算上的讨论，请参阅 §2.1-2。

3.2-4. 基于求积的方法．在第二章问题 8 中讨论的单步求积方法对方程组无需改变而被保留下来，我们只限于所列举的公式，而把误差讨论推迟到 §3.3.

基本求积公式（对于一个纯量函数）由 (2-109) 给出．象单个方程情形一样，我们需要一个方法来预估 $y(x)$ 在点 $x_n + p_k h$ 上的值．我们假设这个预估借助于由增量函数 $\bar{\boldsymbol{\Phi}}_k(\mathbf{y}; p_k h)$ 所定义的方法．（下标 k 表示对每一个横坐标可使用不同预估公式）于是求积方法规定为

$$\boldsymbol{\Phi}(\mathbf{y}; h) = \sum_{k=1}^{\nu} w_k \mathbf{f}(\mathbf{y} + p_k h \bar{\boldsymbol{\Phi}}_{(k)}(\mathbf{y}; p_k h)). \quad (3\text{-}52)$$

3.3. 单步方法的离散误差

本节大部分定理的证明完全相似于 §2.2 中类似定理的证明，于是或把它们省略或给出简述．

3.3-1. 收敛性与相容性．我们总是假设向量函数 $\mathbf{f}(x, \mathbf{y})$ 满足存在性定理 3.1 的条件，那么对于任意初值向量 $\boldsymbol{\eta}$，初值问题

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(a) = \boldsymbol{\eta} \quad (3\text{-}53)$$

在区间 $[a, b]$ 上有一个解．我们还假设对于

$$x \in [a, b], \quad y^i \in (-\infty, +\infty)$$

以及对于使得 $x + h \in [a, b]$ 的一切 $h \geq 0$，在任意给定点 $(x, \mathbf{y})$ 上增量函数 $\boldsymbol{\Phi}(x, \mathbf{y}; h)$ 是确定的．那么对于任意 $h \geq 0$，由 (3-30) 就可能计算出向量 $\mathbf{y}_n$ 只要 $x_n \in [a, b]$．

由增量函数 $\boldsymbol{\Phi}(x, \mathbf{y}; h)$ 所定义的方法便说是收敛的，如果对于任意 $\boldsymbol{\eta}$ 和任意 $x \in [a, b]$，有

$$\lim_{\substack{h \to 0 \\ x_n = x}} \mathbf{y}_n = \mathbf{y}(x). \quad (3\text{-}54)$$

如果

$$\boldsymbol{\Phi}(x,\mathbf{y};0)=\mathbf{f}(x,\mathbf{y}) \tag{3-55}$$

对于 x 及 $\mathbf{y}$ 恒成立，则说此方法对微分方程(3-53)是相容的。

正如 §2.2-1 一样，可以证明以下定理：

定理 3.2. 令函数 $\boldsymbol{\Phi}$ 在上述区域内是连续的(它是 $s+2$ 个变元的函数)，并且存在一个常数 L，使得：对于在那个区域内使 $h\leqslant h_0$ 的一切点 $(x,\mathbf{y}^*;h)$ 和 $(x,\mathbf{y};h)$，其中 $h_0>0$ 是固定常数，都有

$$\|\boldsymbol{\Phi}(x,\mathbf{y}^*;h)-\boldsymbol{\Phi}(x,\mathbf{y};h)\|\leqslant L\|\mathbf{y}^*-\mathbf{y}\| \tag{3-56}$$

成立。那么由 $\boldsymbol{\Phi}$ 所确定的这个方法的收敛的必要与充分条件是这个方法为相容的。

3.3-2. 一个先验界. 令 $\mathbf{y}(x)$ 是初值问题的解，并且

$$\boldsymbol{\Delta}(x,\mathbf{y};h)$$

是由(3-29)确定的。对于微分方程组而言，以下定理综合了定理 2.2 与 2.3 的特点。

定理 3.3. 令 $\boldsymbol{\Phi}(x,\mathbf{y};h)$ 满足定理 3.2 的条件，并且存在常数 $N\geqslant 0$，$p\geqslant 0$ 和 $h_0>0$，使得

$$\|\boldsymbol{\Phi}(x,\mathbf{y}(x);h)-\boldsymbol{\Delta}(x,\mathbf{y}(x);h)\|\leqslant Nh^p,$$
$$x\in[a,b],h\leqslant h_0. \tag{3-57}$$

令 $\{\mathbf{y}_n\}$ 是满足

$$\mathbf{y}_0=\boldsymbol{\eta}, \tag{3-58}$$
$$\mathbf{y}_{n+1}=\mathbf{y}_n+h[\boldsymbol{\Phi}(x_n,\mathbf{y}_n;h)+h^qK\boldsymbol{\theta}_n],$$
$$n=0,1,2,\cdots;x_n\in[a,b]$$

的任意向量序列，其中 $K\geqslant 0$ 和 $q\geqslant 0$ 都是常数，并且向量 $\boldsymbol{\theta}_n$ 满足 $\|\boldsymbol{\theta}_n\|\leqslant 1$，那么，对于 $x_n\in[a,b]$ 及 $h\leqslant h_0$，有

$$\|\mathbf{y}_n-\mathbf{y}(x_n)\|\leqslant h^rN_1E_L(x_n-a), \tag{3-59}$$

其中 $r=\min(p,q)$ 和 $N_1=Nh_0^{p-r}+Kh_0^{q-r}$。

这个定理给我们提供了一个关于误差 $\mathbf{e}_n = \mathbf{y}_n - \mathbf{y}(x_n)$ 的界，即使关系式 (3-30) 仅是近似地满足。

证. 由 (3-58) 减去关系式

$$\mathbf{y}(x_{n+1}) = y(x_n) + h\mathbf{\Delta}(x_n, \mathbf{y}(x_n); h),$$

我们得到

$$\mathbf{e}_{n+1} = \mathbf{e}_n + h[\boldsymbol{\Phi}(x_n, \mathbf{y}_n; h) - \mathbf{\Delta}(x_n, \mathbf{y}(x_n); h) + h^q K \boldsymbol{\theta}_n]. \tag{3-60}$$

应用三角不等式，并利用 (3-56) 及 (3-57)，有

$$\begin{aligned}
&\|\boldsymbol{\Phi}(x_n, \mathbf{y}_n; h) - \mathbf{\Delta}(x_n, \mathbf{y}(x_n); h) + h^q K\boldsymbol{\theta}_n\| \\
&\quad = \|\boldsymbol{\Phi}(x_n, \mathbf{y}_n; h) - \boldsymbol{\Phi}(x_n, \mathbf{y}(x_n); h) \\
&\qquad + \boldsymbol{\Phi}(x_n, \mathbf{y}(x_n); h) - \mathbf{\Delta}(x_n, \mathbf{y}(x_n); h) + h^q K\boldsymbol{\theta}_n\| \\
&\quad \leqslant L\|\mathbf{e}_n\| + h^p N + h^q K.
\end{aligned}$$

从而由 (3-60)，有

$$\|\mathbf{e}_{n+1}\| \leqslant (1 + hL)\|\mathbf{e}_n\| + h^{p+1}N + h^{q+1}K,$$

并令

$$\xi_n = \|\mathbf{e}_n\|, A = 1 + hL, B = h^{p+1}N + h^{q+1}K, \xi_0 = 0.$$

从引理 1.2 便导出结果 (3-59). 不难证明，对于由 (2-51) 确定的不规则点 x，界 (3-59) 仍然成立，如果把 (3-57) 换成

$$\|\boldsymbol{\Phi}(x, \mathbf{y}(x); h\theta(x)) - \mathbf{\Delta}(x, \mathbf{y}(x); h\theta(x))\| \leqslant h^p N. \tag{3-61}$$

3.3-3. 对特殊方法的应用. 关于 L 的值. 我们在本节和下节中将假设函数 $\mathbf{f}$ 不明显地依赖于 x. 正如在 §3.2-1 中所见，这个假设是不失一般性的.

我们假设函数 $\mathbf{f}_j^{(p)}(\mathbf{y})(j = 1, \cdots, s)$ 对于

$$y^i \in (-\infty, +\infty) \ (i = 1, \cdots, s)$$

有界，并令

$$L_p = \sup_{\substack{j=1,\cdots,s \\ -\infty < y^i < \infty}} \|\mathbf{f}_j^{(p)}(\mathbf{y})\|, \quad p = 0, 1, 2, \cdots.$$

基于 Taylor 展式方法(3-38),利用中值定理,我们求得

$$\Phi^i(\mathbf{y}^*;h)-\Phi^i(\mathbf{y};h)$$
$$=\left[f_j^i(y^{[0]})+\frac{h}{2}f_j^{(1)i}(\mathbf{y}^{[1]})+\cdots+\frac{h^{p-1}}{p!}f_j^{(p-1)i}(\mathbf{y})^{[p-1]}\right]$$
$$\times(y^{*j}-y^j),$$

其中 $\mathbf{y}^{[0]},\mathbf{y}^{[1]},\cdots$ 表示某些向量,其分量位于 $\mathbf{y}^*$ 和 $\mathbf{y}$ 的分量之间. 从而推出(3-56)成立,取

$$L=L_0+\frac{h}{2}L_1+\cdots+\frac{h^{p-1}}{p!}L_{p-1}. \tag{3-62}$$

为了确定简化 Runge-Kutta 方法的常数 L,对任意两个向量 $\mathbf{y}^*$ 和 $\mathbf{y}$,我们利用关系式

$$\|\mathbf{f}(\mathbf{y}^*)-\mathbf{f}(\mathbf{y})\|\leqslant L_0\|\mathbf{y}^*-\mathbf{y}\|,$$

推出

$$\left\|\mathbf{f}\left(\mathbf{y}^*+\frac{h}{2\alpha}\mathbf{f}(\mathbf{y}^*)\right)-\mathbf{f}\left(\mathbf{y}+\frac{h}{2\alpha}\mathbf{f}(\mathbf{y})\right)\right\|$$
$$\leqslant L_0\left[\|\mathbf{y}^*-\mathbf{y}\|+\frac{h}{2|\alpha|}\|\mathbf{f}(\mathbf{y}^*)-\mathbf{f}(\mathbf{y})\|\right].$$
$$\leqslant L_0\left(1+\frac{h}{2|\alpha|}L_0\right)\|\mathbf{y}^*-\mathbf{y}\|$$

这便推得(3-56)成立,取

$$L=L_0\left(|1-\alpha|+|\alpha|+\frac{h_0L_0}{2}\right). \tag{3-63}$$

对于经典 Runge-Kutta 方法,可类似分析. 令[1]

$$\mathbf{k}_i=\mathbf{k}_i(\mathbf{y})\quad(i=1,2,3,4)$$

表示向量 $\mathbf{k}_i$ 依赖于 $\mathbf{y}$,我们有

$$\|\mathbf{k}_i(\mathbf{y}^*)-\mathbf{k}_i(\mathbf{y})\|\leqslant L_0\|\mathbf{y}^*-\mathbf{y}\|$$

---

1) 这里下标不表示导数.

一般地，对于 $i=2,3,4$，

$$\begin{aligned}
&\|\mathbf{k}_i(\mathbf{y}^*)-\mathbf{k}_i(\mathbf{y})\| \\
&\qquad \leqslant \|\mathbf{f}(\mathbf{y}^*+hp_i\mathbf{k}_{i-1}(\mathbf{y}^*))-\mathbf{f}(\mathbf{y}+hp_i\mathbf{k}_{i-1}(\mathbf{y})\| \\
&\qquad \leqslant L_0[\|\mathbf{y}^*-\mathbf{y}\|+hp_i\|\mathbf{k}_{i-1}(\mathbf{y}^*)-\mathbf{k}_{i-1}(\mathbf{y})\|],
\end{aligned}$$

其中 $p_2=p_3=\frac{1}{2}$，$p_4=1$．对于 $h\leqslant h_0$，导出

$$\|\mathbf{k}_2(\mathbf{y}^*)-\mathbf{k}_2(\mathbf{y})\|\leqslant L_0\left(1+\frac{h_0L_0}{2}\right)\|\mathbf{y}^*-\mathbf{y}\|,$$

$$\|\mathbf{k}_3(\mathbf{y}^*)-\mathbf{k}_3(\mathbf{y})\|\leqslant L_0\left[1+\frac{h_0L_0}{2}\left(1+\frac{h_0L_0}{2}\right)\right]\|\mathbf{y}^*-\mathbf{y}\|,$$

$$\begin{aligned}
\|\mathbf{k}_4(\mathbf{y}^*)-\mathbf{k}_4(\mathbf{y})\|\leqslant L_0\left\{1+L_0h_0\left[1+\frac{h_0L_0}{2}\left(1+\frac{h_0L_0}{2}\right)\right]\right\} \\
\times\|\mathbf{y}^*-\mathbf{y}\|.
\end{aligned}$$

综合这些结果，我们求得

$$\begin{aligned}
&\|\boldsymbol{\Phi}(\mathbf{y}^*;h)-\boldsymbol{\Phi}(\mathbf{y};h)\| \\
&\quad \leqslant \frac{1}{6}\{\|\mathbf{k}_1(\mathbf{y}^*)-\mathbf{k}_1(\mathbf{y})\|+2\|\mathbf{k}_2(\mathbf{y}^*)-\mathbf{k}_2(\mathbf{y})\| \\
&\qquad +2\|\mathbf{k}_3(\mathbf{y}^*)-\mathbf{k}_3(\mathbf{y})\|+\|\mathbf{k}_4(\mathbf{y}^*)-\mathbf{k}_4(\mathbf{y})\|\} \\
&\quad \leqslant L_0\left[1+\frac{h_0L_0}{2}+\frac{(h_0L_0)^2}{6}+\frac{(h_0L_0)^3}{24}\right] \\
&\qquad \times\|\mathbf{y}^*-\mathbf{y}\|. \qquad (3\text{-}64)
\end{aligned}$$

最后的结果便给出（3-56）中常数 L 的值．这个界与稍大的界

$$L=\frac{e^{h_0L_0}-1}{h_0} \qquad (3\text{-}65)$$

仅相差 $O(h_0^4)$．

对于基本求积方法，如果 $\bar{L}$ 表示对预估方法 $\bar{\boldsymbol{\Phi}}_{(k)}$ 共同的 Lipschitz 常数，我们有

$$\|\boldsymbol{\Phi}(\mathbf{y}^*;h)-\boldsymbol{\Phi}(\mathbf{y};h)\| \leqslant \sum_{k=1}^{\nu} |w_k| \|\mathbf{f}(\mathbf{y}^*+p_k h\bar{\boldsymbol{\Phi}}_{(k)}(\mathbf{y}^*;p_k h))$$
$$- \mathbf{f}(\mathbf{y} + p_k h\bar{\boldsymbol{\Phi}}_{(k)}(\mathbf{y};p_k h))\|$$
$$\leqslant L_0 \sum_{k=1}^{\nu} |w_k|(1 + p_k h\bar{L})\|\mathbf{y}^* - \mathbf{y}\|.$$

如果 $w_k \geqslant 0(k=1,\cdots,\nu)$，我们有 $\Sigma|w_k|p_k = \frac{1}{2}$，于是

$$L = L_0\left(1 + \frac{h_0}{2}\bar{L}\right). \tag{3-66}$$

3.3-4. 对特殊方法的应用；N 的值. 从定理 3.1 的证明便知，初值问题 (3-53) 的精确解 $y(x)$ 满足

$$\|\mathbf{y}(x)\| \leqslant Y,$$

其中 Y 由 (3-15) 确定. 如果由 $\boldsymbol{\Phi}$ 定义的方法为 p 阶，并且 $\boldsymbol{\Phi}$ 和 $\mathbf{f}$ 都是 p 次连续可微，由

$$\Phi^i(x,\mathbf{y};h) - \Delta^i(x,\mathbf{y};h)$$
$$= h^p\left\{\frac{1}{p!}\frac{\partial^p \Phi^i}{\partial h^p}(x,\mathbf{y};h^+)\right.$$
$$\left. - \frac{1}{(p+1)!}f^{(p)i}(x+h^+,\mathbf{z}(x+h^+))\right\},$$

其中 $\mathbf{z}(t)$ 表示通过 $(x,\mathbf{y})$ 的解以及 $0 < h^+ < h$，便推出 (3-57) 成立，取

$$N = \frac{1}{(p+1)!}M_p + \frac{1}{p!}\max_{\substack{\|\mathbf{y}\|\leqslant Y \\ h\leqslant h_0}}\left\|\frac{\partial^p \boldsymbol{\Phi}}{\partial h^p}(\mathbf{y};h)\right\|, \tag{3-67}$$

其中

$$M_p = \max_{\|\mathbf{y}\|\leqslant Y}\|\mathbf{f}^{(p)}(\mathbf{y})\|.$$

对于 Taylor 展式方法 (3-38)，$\partial^p\boldsymbol{\Phi}/\partial h^p = 0$，于是

$$N = \frac{1}{(p+1)!}M_p. \tag{3-68}$$

为了对 Runge-Kutta 型方法写出N的界，我们令

$$D_p = \max_{\substack{1\le j_1,\cdots,j_p\le s \\ \|\mathbf{y}\|\le Y}} \|\mathbf{f}_{j_1 j_2 \cdots j_p}(\mathbf{y})\|.$$

对于简化 Runge-Kutta 方法 (3–47)，利用 (3–43)，有

$$\frac{\partial^2 \Phi}{\partial h^2}(\mathbf{y}; h) = \frac{1}{4\alpha} \mathbf{f}_{jk}\left(\mathbf{y} + \frac{h}{2\alpha}\mathbf{f}(\mathbf{y})\right) f^j(\mathbf{y}) f^k(\mathbf{y}).$$

从而

$$\left\|\frac{\partial^2 \Phi}{\partial h^2}(\mathbf{y}; h)\right\| \le \frac{1}{4|\alpha|} D_2 M_0^2.$$

于是对于这个方法，有

$$N = \frac{1}{6} M_2 + \frac{1}{8|\alpha|} D_2 M_0^2. \tag{3-69}$$

经典 Runge-Kutta 方法 (3–48) 是使得 (3–57) 对 $p = 4$ 是正确的。通过以下辅助性的讨论来求 $\|\partial^4\Phi/\partial h^4\|$ 界的任务便可简化。令 τ 是一个常数，并令 $\mathbf{k} = \mathbf{k}(h)$ 是向量，其分量依赖于 h。我们将计算函数

$$\mathbf{g}(h) = \mathbf{f}(\mathbf{y} + \tau h\,\mathbf{k}(h)) \tag{3-70}$$

的前四阶导数。令

$$\frac{d\mathbf{g}}{dh} = \dot{\mathbf{g}}, \quad \frac{d^2\mathbf{g}}{dh^2} = \mathbf{g}^{[2]}, \cdots$$

并且对依赖于 h 的其它函数使用类似记号，我们求得

$$\dot{\mathbf{g}}(h) = \tau(h k^j)^{\cdot} \mathbf{f}_j, \tag{3-71a}$$

$$\mathbf{g}^{[2]}(h) = \tau^2(hk^j)^{\cdot}(hk^m)^{\cdot}\mathbf{f}_{jm} + \tau(hk^j)^{[2]}\mathbf{f}_j, \tag{3-71b}$$

$$\mathbf{g}^{[3]}(h) = \tau^3(hk^j)^{\cdot}(hk^m)^{\cdot}(hk^n)^{\cdot}\mathbf{f}_{jmn} + 3\tau^2(hk^j)^{[2]}(hk^m)^{\cdot}\mathbf{f}_{jm} + \tau(hk^j)^{[3]}\mathbf{f}_j, \tag{3-71c}$$

$$\begin{aligned}\mathbf{g}^{[4]}(h) = {} & \tau^4(hk^j)^{\cdot}(hk^m)^{\cdot}(hk^n)^{\cdot}(hk^r)^{\cdot}\mathbf{f}_{jmnr} \\ & + 6\tau^3(hk^j)^{[2]}(hk^m)^{\cdot}(hk^n)^{\cdot}\mathbf{f}_{jmn} + 4\tau^2(hk^j)^{[3]}(hk^m)^{\cdot}\mathbf{f}_{jm} \\ & + 3\tau^2(hk^j)^{[2]}(hk^m)^{[2]}\mathbf{f}_{jm} + \tau(hk^j)^{[4]}\mathbf{f}_j,\end{aligned} \tag{3-71d}$$

函数 $\mathbf{f}$ 及其导数中的变元为 $\mathbf{y} + \tau h\mathbf{k}$。

确定经典 Runge-Kutta 方法的关系式(3-51)可写成

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y} + p_i h\, \mathbf{k}_{i-1}), \quad i = 1,2,3,4, \qquad (3-72)$$

其中 $p_1 = 0$，$p_2 = p_3 = \frac{1}{2}$，$p_4 = 1$(向量 $\mathbf{k}_0$ 无需确定)。这些关系式恰好就是形式(3-70)，并且公式(3-71)能使我们表达出 $\mathbf{k}_i$ 的导数.

首先，因为 $\mathbf{f}(\mathbf{y})$ 与 h 无关，故

$$\dot{\mathbf{k}}_1 = \mathbf{k}_1^{[2]} = \mathbf{k}_1^{[3]} = \mathbf{k}_1^{[4]} = \mathbf{0}.$$

因而，从(3-71)，

$$\begin{aligned}
\dot{\mathbf{k}}_2 &= \frac{1}{2} k_1^j \mathbf{f}_j, \\
\mathbf{k}_2^{[2]} &= \frac{1}{4} k_1^j k_1^m \mathbf{f}_{jm}, \\
\mathbf{k}_2^{[3]} &= \frac{1}{8} k_1^j k_1^m k_1^n \mathbf{f}_{jmn}, \\
\mathbf{k}_2^{[4]} &= \frac{1}{16} k_1^j k_1^m k_1^n k_1^r \mathbf{f}_{jmnr}.
\end{aligned} \qquad (3-73)$$

我们的目的是求 $\mathbf{k}_i$ 的导数的界. 为了用最简便方式写出这些界，我们定义一个新的常量 K，使得

$$D_p \leqslant \frac{K}{M^{p-1}}, \quad p = 0,1,\cdots,4, \qquad (3-74)$$

其中 $M = M_0$. 于是 $K \geqslant 1$. 由于 $\|\mathbf{k}_1\| \leqslant M$，我们从(3-73)得到

$$\|\mathbf{k}_2^{[\nu]}\| \leqslant 2^{-\nu} K M, \quad \nu = 1,2,3,4.$$

于是导出

$$\|(h\mathbf{k}_2)^{\cdot}\| \leqslant \|\mathbf{k}_2\| + \|h\dot{\mathbf{k}}_2\| \leqslant M\left[1 + \frac{1}{2}(hK)\right] \leqslant M e^{\theta},$$

其中 $\theta = \frac{1}{2}(hK)$，并且对于 $\nu = 2,3,4$，

$$\|(h\mathbf{k}_2)^{[\nu]}\| \leqslant \|\nu \mathbf{k}_2^{[\nu-1]}\| + \|h\mathbf{k}_2^{[\nu]}\| \leqslant 2^{-\nu+1}\nu K M e^{\theta}.$$

在这里我们用 e^{ah} 来代替形式为 $1+ah$ 的因子，因为这就可能确切地估计出这个因子乘积的上界如下：

$$\Pi(1+a_i h)\leqslant e^{h\Sigma a_i}.$$

从 (3-71)，令 $\mathbf{g}(h)=\mathbf{k}_3$，$\mathbf{k}(h)=\mathbf{k}_2$，现在我们容易求得

$$\|\dot{\mathbf{k}}_3\|\leqslant\frac{1}{2}KMe^{\theta},$$

$$\|\mathbf{k}_3^{[2]}\|\leqslant\frac{1}{4}KM(1+2K)e^{2\theta},$$

$$\|\mathbf{k}_3^{[3]}\|\leqslant\frac{1}{8}KM(1+9K)e^{3\theta},$$

$$\|\mathbf{k}_3^{[4]}\|\leqslant\frac{1}{16}KM(1+28K+12K^2)e^{4\theta},$$

由此我们得到

$$\|(h\mathbf{k}_3)^{\cdot}\|\leqslant\|\dot{\mathbf{k}}_3\|+\|h\dot{\mathbf{k}}_3\|\leqslant M\left[1+\frac{1}{2}(hK)e^{\theta}\right]$$

$$\leqslant Me^{2\theta'},$$

其中

$$\theta'=\frac{1}{4}(hK)e^{\theta}=\frac{1}{4}(hK)e^{\frac{1}{2}(hK)}. \qquad (3\text{-}75)$$

用类似方法，利用 $\theta\leqslant 2\theta'$，我们求得

$$\|(h\mathbf{k}_3)^{[2]}\|\leqslant KMe^{5\theta'},$$

$$\|(h\mathbf{k}_3)^{[3]}\|\leqslant\frac{3}{4}KM(1+2K)e^{8\theta'},$$

$$\|(h\mathbf{k}_3)^{[4]}\|\leqslant\frac{1}{2}KM(1+9K)e^{9\theta'}.$$

利用 (3-71d)，取 $\mathbf{g}(h)=\mathbf{k}_4$，$\mathbf{k}(h)=\mathbf{k}_3$，用同样方法，我们得到

$$\|k_4^{[4]}\|\leqslant\frac{1}{2}MK(2+19K+27K^2)e^{9\theta'}.$$

由于

$$\frac{\partial^4\Phi}{\partial h^4}(\mathbf{y};h)=\frac{1}{3}\mathbf{k}_2^{[4]}+\frac{1}{3}\mathbf{k}_3^{[4]}+\frac{1}{6}\mathbf{k}_4^{[4]}.$$

综合这些结果，得到

$$\left\|\frac{\partial^4\Phi}{\partial h^4}(y;h)\right\|\leqslant\frac{1}{24}MK(5+52K+60K^2)e^{9\theta'},\quad(3-76)$$

其中M和K由(3-74)确定，而θ'由(3-75)确定。从下面给出的公式(3-84)推出M_4的界可用K和M表示如下：

$$M_4=\max_{\|\mathbf{y}\|\leqslant Y}\|\mathbf{f}^{[4]}(\mathbf{y})\|\leqslant MK(1+11K+7K^2+K^3).\quad(3-77)$$

现在我们便可应用(3-67)。由于$\theta'=O(h)$，$e^{9\theta'}=1+O(h)$，并且如果结合(3-76)和(3-77)的二个界，

$$N=KM\frac{49+524K+468K^2+24K^3}{2880}e^{9\theta'},\quad(3-78)$$

这并没有多大影响。这就是对于 Runge-Kutta 方法所需要的常数N的界。

为了用求积方法来估计N，我们用$\bar{N}$及$\bar{p}$表示对应于用(3-52)中增量函数$\bar{\Phi}$定义的预估方法的常数N和p的最大和最小值。用$\mathbf{z}(t)$表示通过$(x,\mathbf{y})$的$\mathbf{z}'=\mathbf{f}(t,\mathbf{z})$的解，那么就可估计$h\bar{\Phi}(x,\mathbf{y},h)-h\Delta(x,\mathbf{y};h)$如下：

$$\begin{aligned}&\left\|h\sum_{k=1}^{\nu}w_k\mathbf{f}(x+p_kh,\mathbf{y}+p_kh\bar{\Phi}_{(k)}(x,\mathbf{y};p_kh)-\mathbf{z}(x+h)+\mathbf{y}\right\|\\&\leqslant\left\|h\sum_{k=1}^{\nu}w_k\mathbf{f}(x+p_kh,\mathbf{z}(x+p_kh)-\mathbf{z}(x+h)+\mathbf{y}\right\|\\&\quad+Lh\sum_{k=1}^{\nu}|w_k|\|p_kh\bar{\Phi}_{(k)}(x,y;p_kh)-[\mathbf{z}(x+p_kh)-\mathbf{y}]\|\\&\leqslant\left\|h\sum_{k=1}^{\nu}w_k\mathbf{z}'(x+p_kh)-[\mathbf{z}(x+h)-\mathbf{y}]\right\|\\&\quad+L\bar{N}h\sum_{k=1}^{\nu}|w_k|(p_kh)^{\bar{p}+1}.\end{aligned}$$

第一项以 $|C|h^{\mu+1}M_\mu$ 为界。如果 $w_k \geq 0$ 和 $C > 0$，我们有

$$\sum_{k=1}^{\nu} |w_k| p_k^{\bar{p}+1} \leq \int_0^1 t^{\bar{p}+1} dt = \frac{1}{\bar{p}+2}.$$

从而第二项估计为

$$h^{\bar{p}+2} L \bar{N} \frac{1}{\bar{p}+2}.$$

于是 $p = \min(\bar{p}+1, \mu)$ 且 (3-57) 成立，取

$$N = \frac{L\bar{N}}{\bar{p}+2} h_0^{\bar{p}+1-p} + CM_\mu h_0^{\mu-\bar{p}}. \qquad (3\text{-}79)$$

因此仅当 $\bar{p}+1 \geq \mu$ 时使用求积公式优越性为最大。

3.3-5. 主误差函数. 如果 N 表示由 (3-57) 确定且在前节已经估计的常数，那么量 $h^{p+1}N$ 表示局部离散误差的界。我们的经验是这样的界大大超过真正误差，并且由误差渐近公式可给出误差真正大小的较准表示。如果由 $\boldsymbol{\Phi}(x,\mathbf{y};h)$ 规定的方法为 p 阶精确，并且 $\boldsymbol{\Phi}(x,\mathbf{y};h)$ 和 $\boldsymbol{\Delta}(x,\mathbf{y};h)$ 有 $p+1$ 阶连续导数，我们可写成

$$\boldsymbol{\Phi}(x,\mathbf{y};h) - \boldsymbol{\Delta}(x,\mathbf{y};h) = h^p \boldsymbol{\varphi}(x,\mathbf{y}) + O(h^{p+1}), \qquad (3\text{-}80)$$

其中向量 $\boldsymbol{\varphi}$ 仍依赖于 x 和 $\mathbf{y}$ 且恒不为零.

象一维情形一样，函数 $\boldsymbol{\varphi}(x,\mathbf{y})$ 称为方法的主误差函数. 量 $h^p\boldsymbol{\varphi}(x,\mathbf{y})$ 和 $h^{p+1}\boldsymbol{\varphi}(x,\mathbf{y})$ 是在点 $(x,\mathbf{y})$ 上方法的相对和绝对局部截断误差的近似.

利用公式，由 (3-80) 导出

$$\boldsymbol{\varphi}(x,\mathbf{y}) = \frac{1}{p!}\frac{\partial^p \boldsymbol{\Phi}}{\partial h^p}(x,\mathbf{y};0) - \frac{1}{(p+1)!}\mathbf{f}^{(p)}(x,\mathbf{y}). \qquad (3\text{-}81)$$

我们对以前所讨论的一些特殊方法来确定 $\boldsymbol{\varphi}$. 不失一般性，还是假设 $\mathbf{f}$, $\boldsymbol{\Phi}\boldsymbol{\Delta}$ 及 $\boldsymbol{\varphi}$ 都不显式地依赖于 x。

对于 Taylor 展式方法，我们立得

$$\boldsymbol{\varphi}(\mathbf{y}) = -\frac{1}{(p+1)!}\mathbf{f}^{(p)}(\mathbf{y}), \qquad (3\text{-}82)$$

因为$\boldsymbol{\Phi}$是 h 的 $p-1$ 次多项式.

对于由 (3–47) 确定的简化 Runge-Kutta 方法的情形，我们求得

$$\frac{1}{2}\frac{\partial^2\boldsymbol{\Phi}}{\partial h^2}(\mathbf{y};0) = \frac{1}{8\alpha}\mathbf{f}_{jk}f^kf^j,$$

函数 $\mathbf{f}$ 及其导数中的变元均为 $\mathbf{y}$. 我们从这个公式减去

$$\frac{1}{6}\mathbf{f}''(\mathbf{y}) = \frac{1}{6}\mathbf{f}_{jk}f^jf^k + \frac{1}{6}\mathbf{f}_j f^j_k f^k,$$

因为 $\mathbf{f}' = \mathbf{f}_k f^k$. 其结果可写成如下形式:

$$\boldsymbol{\varphi}(\mathbf{y}) = \left(\frac{1}{8\alpha} - \frac{1}{6}\right)\mathbf{f}'' - \frac{1}{8\alpha}\mathbf{f}_j f'^j. \qquad (3\text{-}83)$$

如果我们选取 $\alpha = \frac{3}{4}$，于是出现主误差函数具有特别简单的形式.

经典 Runge-Kutta 方法可作类似地处理. 我们仅引用经过冗长计算所得的结果. 结合在 § 3.2–2 中确定的向量$\mathbf{A}$, $\mathbf{B},\cdots,\mathbf{H}$，我们令

$$\begin{aligned}
&\mathbf{I} = \mathbf{f}_{jklm}f^jf^kf^lf^m, && \mathbf{N} = \mathbf{f}_m f^m_{jl} f^l_k f^j f^k,\\
&\mathbf{J} = \mathbf{f}_{jkm}f^m_l f^j f^k f^l, && \mathbf{P} = \mathbf{f}_l f^l_m f^m_{jk} f^j f^k,\\
&\mathbf{K} = \mathbf{f}_{jm}f^m_{kl} f^j f^k f^l, && \mathbf{Q} = \mathbf{f}_m f^m_l f^l_k f^k_j f^j,\\
&\mathbf{L} = \mathbf{f}_{jm}f^m_l f^l_k f^j f^k, && \mathbf{R} = \mathbf{f}_{lm} f^m_k f^l_j f^j f^k,\\
&\mathbf{M} = \mathbf{f}_j f^j_{klm} f^k f^l f^m,
\end{aligned}$$

其变元均理解为 $\mathbf{y}$，可以证明

$$\frac{1}{4!}\frac{\partial^4\boldsymbol{\Phi}}{\partial h^4}(\mathbf{y};0) = \frac{1}{576}\{5\mathbf{I} + 30\mathbf{J} + 18\mathbf{K} + 24\mathbf{L} + 4\mathbf{M} + 12\mathbf{N} + 6\mathbf{P} + 18\mathbf{R}\},$$

另一方面，

$$\frac{1}{5!}\mathbf{f}^{\mathrm{V}}(y)=\frac{1}{120}\{\mathbf{I}+6\mathbf{J}+4\mathbf{K}+\mathbf{M}+3\mathbf{N}+\mathbf{P}+4\mathbf{L}+\mathbf{Q}+3\mathbf{R}\}. \tag{3-84}$$

于是导出

$$\boldsymbol{\varphi}(\mathbf{y})=\frac{1}{2880}\{\mathbf{I}+6\mathbf{J}-6\mathbf{K}+24\mathbf{L}-4\mathbf{M}-12\mathbf{N}+6\mathbf{P}-24\mathbf{Q}+18\mathbf{R}\}. \tag{3-85}$$

尽可能用 $\mathbf{f}^{(k)}$ 的分量来表达 $\mathbf{f}$ 的累次导数，便可简化这个表达式，其结果是

$$\boldsymbol{\varphi}(\mathbf{y})=\frac{1}{2880}\mathbf{f}^{\mathrm{IV}}-\frac{1}{576}\mathbf{f}_j f'''^j+\frac{1}{288}[\mathbf{f}_j f^j_k-\mathbf{f}_{jk}f^j]f''^k+\frac{1}{192}[2\mathbf{f}_{jk}f^j f^k_l-2\mathbf{f}_{jl}f^j_k f^k_l+\mathbf{f}_{jl}f^j_k f^k]f'^l. \tag{3-86}$$

对于 $s=2$，$y^1=x$，$y^2=y$，$f^1=1$，$f^2=f(x,y)$，向量(3-86)的第二个分量给出(2-36)。

为了讨论基于求积方法的主误差函数，令

$$\mathscr{A}(\mathbf{y})=(a^{ij}(\mathbf{y}))$$

表示 $s\times s$ 矩阵，其元素为

$$a^{ij}(\mathbf{y})=f^i_j(\mathbf{y}),\quad i,j=1,\cdots,s. \tag{3-87}$$

如果 $\mathbf{k}$ 是任意向量，那么我们可写成

$$\mathbf{f}(\mathbf{y}+h\mathbf{k})-\mathbf{f}(\mathbf{y})=h\mathscr{A}(\mathbf{y})\mathbf{k}+O(h^2).$$

现令 $\boldsymbol{\Phi}(\mathbf{y};h)$ 由(3-52)确定，其中用于预估的增量函数 $\bar{\boldsymbol{\Phi}}(\mathbf{y};h)$ 的阶为 $\bar{p}=\mu-1$，并且与其相关的主误差函数为 $\bar{\boldsymbol{\varphi}}(\mathbf{y})$。于是我们有

$$\boldsymbol{\Phi}(\mathbf{y};h)=\sum_{k=1}^{\nu}w_k\mathbf{f}(\mathbf{y}+p_k h\bar{\boldsymbol{\Phi}}(\mathbf{y};p_k h))$$

$$=\sum_{k=1}^{\nu}w_k[\mathbf{f}(\mathbf{y}+p_k h\boldsymbol{\Delta}(\mathbf{y};p_k h))+(p_k h)^{\mu}\mathscr{A}(\mathbf{y})\bar{\boldsymbol{\varphi}}(\mathbf{y})$$

$$+ O(h^{\mu+1})] = \Delta(\mathbf{y};h) - h^{\mu}C\mathbf{f}^{(\mu)}(\mathbf{y})$$

$$+ h^{\mu}\left(\frac{1}{\mu+1} - C\,\mu!\right)\mathscr{A}(\mathbf{y})\bar{\boldsymbol{\varphi}}(\mathbf{y}) + O(h^{\mu+1}),$$

我们断定

$$\boldsymbol{\varphi}(\mathbf{y};h) = -C\,\mathbf{f}^{(\mu)}(\mathbf{y}) + \left[\frac{1}{\mu+1} - C\,\mu!\right]\mathscr{A}(\mathbf{y})\bar{\boldsymbol{\varphi}}(\mathbf{y}). \tag{3-88}$$

如果 $\bar{p}+1>\mu$，则可简化成

$$\boldsymbol{\varphi}(\mathbf{y}) = -C\,\mathbf{f}^{(\mu)}(\mathbf{y}).$$

在这种情形下，主误差函数与真解的导数成比例.

以上给出的 $\boldsymbol{\varphi}(\mathbf{y})$的大部分公式在实际计算时均不实用．因此就需要一个数值方法来近似 $\boldsymbol{\varphi}(\mathbf{y})$．在一维情形，这个方法是由局部外推到极限的方法来提供的，这便是定理 2.5 所阐述的内容．在微分方程组的情形，类似的方法也成立．令 $\boldsymbol{\delta}_1$ 和 $\boldsymbol{\delta}_2$ 表示在长度为 h 的区间上数值解的增量，它们是用由 $\boldsymbol{\Phi}(\mathbf{y};h)$ 确定的方法分别取步长为 h 的一步和步长为 $\frac{1}{2}h$ 的二步计算出来的．采用记号：

$$\boldsymbol{\delta}_1 = h\boldsymbol{\Phi}(\mathbf{y};h),$$

$$\boldsymbol{\delta}_2 = \frac{1}{2}h\left[\boldsymbol{\Phi}\left(\mathbf{y};\frac{1}{2}h\right) + \boldsymbol{\Phi}\left(\mathbf{y}+\frac{1}{2}h\boldsymbol{\Phi}(\mathbf{y});\frac{1}{2}h\right)\right].$$

象 §2.2-7 一样，完全可以证明

$$\boldsymbol{\varphi}(\mathbf{y}) = -\frac{h^{-p-1}}{1+2^{-p}}(\boldsymbol{\delta}_2-\boldsymbol{\delta}_1) + O(h). \tag{3-89}$$

3.3-6. 离散误差的渐近公式．我们还允许函数 $\mathbf{f}(x,\mathbf{y})$和 $\boldsymbol{\Phi}(x,\mathbf{y};h)$ 显式地依赖于 x，但仍用 $\mathbf{f}_i$ 表示向量 $\partial\mathbf{f}/\partial y^i$（然而为了表示对于 h 的导数，便保留下标 h）．假设由 $\boldsymbol{\Phi}(x,\mathbf{y};h)$ 确定的方法 p 阶精确并且满足定理 3.3 的条件．另外，对于 $x\in(a,b)$，$y^i\in(-\infty,\infty)(i=1,\cdots,s)$ 及 $h\leqslant h_0$，$\boldsymbol{\Phi}(x,$

$\mathbf{y}; h$）对所有变元有 $p+2$ 阶连续导数． 于是主误差函数 $\boldsymbol{\varphi}(x,\mathbf{y})$ 存在且连续并有连续的导数． 我们还假设 $\boldsymbol{\Phi}$ 关于 h 和 y^i 的二阶导数有界：

$$\begin{aligned}
&\|\boldsymbol{\Phi}_{hi}(x,\mathbf{y};h)\| \leqslant K_1, \\
&\|\boldsymbol{\Phi}_{ij}(x,\mathbf{y};h)\| \leqslant K_1, \quad i,j=1,\cdots,s, \qquad (3\text{-}90)\\
&x\in[a,b], y^i\in(-\infty,\infty),\ (i=1,\cdots,s) \quad h\leqslant h_0.
\end{aligned}$$

如果 $\mathbf{y}(x)$ 表示初值问题（3-9）的解，并且向量 $\mathbf{y}_n$ 由（3-30）所确定，于是根据定理 3.3，误差 $\mathbf{e}_n=\mathbf{y}_n-\mathbf{y}(x_n)$ 满足

$$\|\mathbf{e}_n\| \leqslant K_2 h^p, \qquad (3\text{-}91)$$

其中 $K_2 = NE_L(b-a)$。

我们再转到离散误差公式（3-60），把它写成如下形式：

$$\begin{aligned}
\mathbf{e}_{n+1} = \mathbf{e}_n + h[&\boldsymbol{\Phi}(x_n,\mathbf{y}_n;h) - \boldsymbol{\Phi}(x_n,\mathbf{y}(x_n);h)\\
&+\boldsymbol{\Phi}(x_n,\mathbf{y}(x_n);h) - \boldsymbol{\Delta}(x_n,\mathbf{y}(x_n);h)]. \qquad (3\text{-}92)
\end{aligned}$$

利用带有余项的 Taylor 公式，有

$$\begin{aligned}
&\boldsymbol{\Phi}(x_n,\mathbf{y}_n;h) - \boldsymbol{\Phi}(x_n,\mathbf{y}(x_n);h)\\
&\quad = \boldsymbol{\Phi}(x_n,\mathbf{y}(x_n)+e_n;h) - \boldsymbol{\Phi}(x_n,\mathbf{y}(x_n);h)\\
&\quad = \boldsymbol{\Phi}_i(x_n,\mathbf{y}(x_n);h)\mathbf{e}_n^i + \frac{1}{2}\theta_n K_1 K_2^2 h^{2p},
\end{aligned}$$

其中 $\|\theta_n\| \leqslant 1$． 如果 $\mathscr{A}(x,\mathbf{y}) = (a^{ij})$ 表示矩阵，其分量为

$$a^{ij} = f_j^i(x,\mathbf{y}),$$

并规定矩阵 $\mathbf{G}(x)$ 为

$$\mathbf{G}(x) = \mathscr{A}(x,\mathbf{y}(x)).$$

按 h 的幂次展开并利用（3-90），则得

$$\boldsymbol{\Phi}_i(x_n,\mathbf{y}(x_n);h) = \mathbf{G}(x_n)\mathbf{e}_n + h^{p+1}\theta_n' K_1 K_2,$$

其中仍有 $\|\theta_n'\| \leqslant 1$． 最后，令

$$K_3 = \frac{1}{(p+1)!}\sup_{\substack{x\in[a,b]\\ h\leqslant h_0}}\left\|\frac{\partial^{p+1}}{\partial h^{p+1}}[\boldsymbol{\Phi}(x,\mathbf{y}(x);h) - \boldsymbol{\Delta}(x,\mathbf{y}(x);h)]\right\|.$$

再利用 Taylor 公式，我们有

$$\boldsymbol{\Phi}(x,\mathbf{y}(x);h)-\boldsymbol{\Delta}(x,\mathbf{y}(x);h)$$
$$=h^{p}\boldsymbol{\varphi}(x,\mathbf{y}(x))+h^{p+1}\boldsymbol{\theta}K_3,$$
其中 $\boldsymbol{\theta}$ 依赖于 x 和 h，并且 $\|\boldsymbol{\theta}\|\leqslant 1$. 规定伸缩误差 $\bar{\mathbf{e}}_n$ 为
$$\bar{\mathbf{e}}_n=h^{-p}\mathbf{e}_n,\quad n=0,1,2,\cdots.$$
从而可把 (3-92) 写成形式
$$\bar{\mathbf{e}}_{n+1}=\bar{\mathbf{e}}_n+h[\mathbf{G}(x_n)\bar{\mathbf{e}}_n+\boldsymbol{\varphi}(x_n,\mathbf{y}(x_n))]$$
$$+\boldsymbol{\theta}_n''h^2[K_1K_2+h_0^{p-1}K_1K_2^2+K_3].$$
这个关系式可看成对函数 $\mathbf{e}(x)$的以下初值问题：
$$\begin{aligned}&\mathbf{e}(a)=\mathbf{0},\\&\mathbf{e}'=\mathbf{G}(x)\mathbf{e}+\boldsymbol{\varphi}(x,\mathbf{y}(x))\end{aligned}\tag{3-93}$$
解的一个 Euler 近似(直到 $O(h^2)$).

利用本节开始时叙述的假设，这个问题的解 $\mathbf{e}(x)$存在并且对于 $x\in[a,b]$ 有连续的二阶导数. 从而我们可应用定理 3.3，并用 $\mathbf{e}(x)$ 代替 $\mathbf{y}(x)$，$\mathbf{G}(x)\mathbf{e}+\boldsymbol{\varphi}(x,\mathbf{y}(x))$ 代替 $\mathbf{f}(x,\mathbf{y})=\boldsymbol{\Phi}(x,\mathbf{y};h)$，以及令
$$\begin{aligned}&p\geqslant 1,\\&L_1=\max_{x\in[a,b]}\left\{\max_{1\leqslant i\leqslant s}\sum_{j=1}^{s}|g^{ij}(x)|\right\},\\&N_1=\frac{1}{2}\max_{x\in[a,b]}\|\mathbf{e}''(x)\|,\\&K_4=K_1K_2(1+h_0^{p-1}K_2)+K_3.\end{aligned}\tag{3-94}$$
于是导出
$$\|\bar{\mathbf{e}}_n-\mathbf{e}(x_n)\|\leqslant hK_5,\quad x_n\in[a,b],$$
其中
$$K_5=(N_1+K_4)E_{L_1}(b-a).\tag{3-95}$$
我们证明了以下结果：

定理 3.4. 在本节开始所述的假设下，利用由 $\boldsymbol{\Phi}(x,\mathbf{y};h)$ 所确定的单步方法，初值问题 (3-9) 的解的离散误差满足
$$\mathbf{e}_n=h^p\mathbf{e}(x_n)+h^{p+1}\boldsymbol{\theta}_nK_5,\tag{3-96}$$

其中 $\mathbf{e}(x)$是 (3-93) 的解，$\|\theta_n\| \leqslant 1$，K_5 由 (3-95) 所确定.

在理论上，通过积分方程组 (3-93) 以及从略去含有 θ_n 的 (3-96) 来近似计算误差 $\mathbf{e}_n$，这个结果可改进给定的数值解. 更为实际的是，定理证实了外推到极限对方程组是正确的. 象在 §2.2-7 那样来规定 $\mathbf{y}(x,h)$，对于 $q \neq 1$，有

$$\mathbf{y}(x) = \frac{\mathbf{y}(x,qh) - q^p\mathbf{y}(x,h)}{1-q^p} + \theta\,\frac{1+q}{1-q^p}\,q^r h^{p+1}K_5, \tag{3-97}$$

其中 $\|\theta\| \leqslant 1$. 令 $\theta = 0$，所得 $\mathbf{y}(x)$ 值其误差仅为 $O(h^{p+1})$.

定理 3.4 容易推广到变步长情形. 如果点 x_n 由 (2-51) 确定，并且用 (3-61) 代替 (3-57)，则 (3-96) 仍然成立. 如果 $\mathbf{e}(x)$规定为

$$\begin{aligned} &\mathbf{e}(a) = \mathbf{0}, \\ &\mathbf{e}' = \mathbf{G}(x)\mathbf{e} + [\theta(x)]^p\boldsymbol{\varphi}(x,\mathbf{y}(x)). \end{aligned} \tag{3-98}$$

3.3-7. 例. 为了方便，我们令(仅用于本节):

$$\begin{aligned} &y^1 = y,\quad e^1 = \eta,\quad \varphi^1 = \varphi, \\ &y^2 = z,\quad e^2 = \zeta,\quad \varphi^2 = \psi. \end{aligned}$$

现在讨论非线性初值问题:

$$\begin{aligned} &y' = -z, \qquad y(0) = a^{-1}, \\ &z' = -2y^3,\quad z(0) = a^{-2} \end{aligned} \tag{3-99}$$

的解，其中 $a > 0$. 采用单步方法 (3-52)，其中基本求积公式为[1)]

$$h^{-1}\int_x^{x+h} f(t)dt = \frac{1}{4}f(x) + \frac{3}{4}f\left(x+\frac{2}{3}h\right) + \frac{1}{216}h^3f'''(\xi), \tag{3-100}$$

并且 $\overline{\Phi}_1$ 是改进 Euler 方法 [(3-47) 中的 $\alpha=1$] 的增量函数.

1) 这是 Radau 求积公式的特殊情形，见 Hildebrand [1956], p. 338.

(3-99) 的精确解为 $y(x)=\xi^{-1}$, $z(x)=\xi^{-2}$, 其中 $\xi=x+a$. 导出

$$\mathscr{A}(\mathbf{y}(x))=\begin{pmatrix}0 & -1\\ -6\xi^{-2} & 0\end{pmatrix}.$$

根据 (3-83), 改进 Euler 方法的主误差向量有分量

$$\bar{\varphi}(x)=\xi^{-4},\quad \bar{\psi}(x)=\frac{5}{2}\xi^{-5}.$$

由于 $\bar{p}=2$, $\mu=3$, 有 $\bar{p}+1=\mu$, 从而 $p=3$. 从(3-88), 我们确定主误差向量的分量如下:

$$\varphi(x)=-\frac{2}{3}\xi^{-5},\ \psi(x)=-\frac{17}{9}\xi^{-6}.$$

于是伸缩误差向量的分量 $\eta(x)$和 $\zeta(x)$满足

$$\begin{aligned}\eta'&=-\zeta-\frac{2}{3}\xi^{-5}, & \eta(0)&=0,\\ \zeta'&=-6\xi^{-2}\eta-\frac{17}{9}\xi^{-6}, & \zeta(0)&=0.\end{aligned}\tag{3-101}$$

微分第一个方程并以第二个方程代入 ζ' 便消去 ζ. 我们积分方程组得到解:

$$\eta(x)=\frac{1}{126a^4}\left\{2\left(\frac{\xi}{a}\right)^3-49\left(\frac{\xi}{a}\right)^{-2}+47\left(\frac{\xi}{a}\right)^{-4}\right\},$$

表 3.1

p	y_n	$y_n-y(6)$	$n^2\eta(6)$	z_n	$z_n-z(6)$	$n^3\zeta(6)$
1	0.1324308	0.0074308	0.0077580	0.0126438	−0.0029812	−0.0030205
2	·1259683	·0009683	·0009698	·0152445	−.0003805	−.0003776
3	·1251219	·0001219	·0001212	·0155774	−.0000476	−.0000472
4	·1250152	·0000152	·0000152	·0156191	−.0000059	−.0000059
5	·1250019	·0000019	·0000019	·0156243	−.0000007	−.0000007
6	·1250002	·0000002	·0000002	.0156249	−.0000001	−.0000001
7	·1250000	·0000000	·0000000	.0156250	−.0000000	−.0000000

$$\zeta(x)=\frac{1}{63a^5}\left\{-3\left(\frac{\xi}{a}\right)^2-49\left(\frac{\xi}{a}\right)^{-3}+52\left(\frac{\xi}{a}\right)^{-5}\right\}. \tag{3-102}$$

在表 3.1 中，我们给出用步长 $h=2^{-p}$，$p=1(1)7$ 对于 $x_n=6$ 的数值解，真正误差以及由 (3-96) 给出的预估误差．

3.4. 用单步方法积分方程组的舍入误差

3.4-1. 累积舍入误差的一个先验界. 如果采用定点运算，数值近似 $\tilde{\mathbf{y}}_n$ 满足把 (3-30) 换成的关系式

$$\tilde{\mathbf{y}}_{n+1}=\tilde{\mathbf{y}}_n+[h\tilde{\boldsymbol{\Phi}}(x_n,\tilde{\mathbf{y}}_n;h)]^*,\quad n=0,1,2,\cdots, \tag{3-103}$$

其中 $\tilde{\boldsymbol{\Phi}}(x_n,\tilde{\mathbf{y}}_n;h)$ 表示对 $\boldsymbol{\Phi}(x_n,\tilde{\mathbf{y}}_n;h)$ 的数值近似（注意这二个函数中的变元是相同的)，并用“*”表示舍入后的值. 从解析角度讨论，更为方便地是假设数值 $\tilde{\mathbf{y}}_n$ 以关系式

$$\tilde{\mathbf{y}}_{n+1}=\tilde{\mathbf{y}}_n+h\tilde{\boldsymbol{\Phi}}(x_n,\tilde{\mathbf{y}}_n;h)+\boldsymbol{\varepsilon}_{n+1} \tag{3-104}$$

相联接，其中 $\boldsymbol{\varepsilon}_{n+1}$ 称为局部舍入误差向量. 在上述情形，$\boldsymbol{\varepsilon}_{n+1}$ 由

$$\boldsymbol{\varepsilon}_{n+1}=[h\tilde{\boldsymbol{\Phi}}(x_n,\tilde{\mathbf{y}}_n;h)]^*-h\boldsymbol{\Phi}(x_n,\tilde{\mathbf{y}}_n;h) \tag{3-105}$$

给出. 从而它是容易估计的. 但是方程 (3-104) 的优点不限于把它应用到定点单倍精度算术运算的特殊情形（见 §3.4-8).

本节的目的是导出累积舍入误差

$$\mathbf{r}_n=\tilde{\mathbf{y}}_n-\mathbf{y}_n$$

的界，在单独假设

$$\|\boldsymbol{\varepsilon}_n\|\leqslant\varepsilon \tag{3-106}$$

之下，其中 ε 是一个常数$\left(\text{常常可假设 }\varepsilon=\frac{1}{2}su\right)$. 从 (3-104)

减去 (3-30) 便推得

$$\mathbf{r}_{n+1} = \mathbf{r}_n + h\{\boldsymbol{\Phi}(x_n, \tilde{\mathbf{y}}_n; h) - \boldsymbol{\Phi}(x_n, \mathbf{y}_n; h)\} + \boldsymbol{\varepsilon}_{n+1}. \tag{3-107}$$

如果 L 表示由 $\boldsymbol{\Phi}$ 所确定的方法的 Lipschitz 常数，利用 (3-106)，有

$$\|\mathbf{r}_{n+1}\| \leqslant \|\mathbf{r}_n\| + hL\|\mathbf{r}_n\| + \varepsilon.$$

应用引理 1.2 及 $\mathbf{r}_0 = 0$ 的事实，采用类似的方式，由这个不等式推出以下定理：

定理 3.5. 如果由增量函数 $\boldsymbol{\Phi}(x, \mathbf{y}; h)$ 确定的方法满足 (3-56)，并且按(3-106)局部舍入误差有界，那么累积舍入误差的界如下：

$$\|\mathbf{r}_n\| \leqslant \frac{\varepsilon}{h} E_L(x_n - a), x_n \in [a, b]. \tag{3-108}$$

这个基本结果是假设舍入误差不仅在每个节点 x_n 上而且在 s 个联立的微分方程中的每个方程之间互相应有系统地增加。因此所得到的估计是很悲观的。从而对局部舍入误差影响的更为实际的估计较之在单个方程的情形更具有迫切性。

3.4-2. 累积舍入误差对局部舍入误差的依赖关系；一个特殊情形. 我们采用 Euler 方法，从讨论线性方程组

$$\mathbf{y}' = \mathbf{G}(x)\mathbf{y} \tag{3-109}$$

开始。$s \times s$ 矩阵 $\mathbf{G}(x)$的元素假设为给定的 x 的连续函数. 在这种情形，Euler 方法的增量函数为

$$\boldsymbol{\Phi}(x, \mathbf{y}; h) = \mathbf{G}(x)\mathbf{y}.$$

从而关系式 (3-107) 有简单形式：

$$\mathbf{r}_{n+1} = \mathbf{r}_n + h\mathbf{G}(x_n)\mathbf{r}_n + \boldsymbol{\varepsilon}_{n+1} \tag{3-110}$$

我们的目的在于把 $\mathbf{r}_n$ 用向量 $\boldsymbol{\varepsilon}_m (m \leqslant n)$ 表示如下：

$$\mathbf{r}_n = \sum_{m=1}^{n} \mathbf{D}_{n,m}\boldsymbol{\varepsilon}_m, \tag{3-111}$$

这里 $\mathbf{D}_{n,m}$ 表示待定的矩阵。把 (3-111) 代入到 (3-110) 中，求得

$$\sum_{m=1}^{n+1} \mathbf{D}_{n+1,m}\boldsymbol{\varepsilon}_m = \sum_{m=1}^{n} (\mathbf{I} + h\mathbf{G}(x_n))\mathbf{D}_{n,m}\boldsymbol{\varepsilon}_m + \boldsymbol{\varepsilon}_{n+1}, \tag{3-112}$$

其中 $\mathbf{I}$ 表示单位矩阵。这个恒等式对所有待定向量 $\boldsymbol{\varepsilon}_n$ 必成立。尤其是，当仅有一个 $\boldsymbol{\varepsilon}_m$ 异于零以及对唯一不为零的 $\boldsymbol{\varepsilon}_m$ 作一切可能的选取时，它一定成立。这就推出在 (3-112) 两边的每个 $\boldsymbol{\varepsilon}_m$ 的乘数必定恒等。于是导出条件：

$$\begin{aligned} &\mathbf{D}_{n+1,n+1} = \mathbf{I}, \quad n = 0,1,2,\cdots \\ &\mathbf{D}_{n+1,m} = \mathbf{D}_{n,m} + h\mathbf{G}(x_n)\mathbf{D}_{n,m}, \\ &m = 1,2,3,\cdots, \quad n = m, \ m+1,\cdots. \end{aligned} \tag{3-113}$$

反之，这些条件则完全确定了矩阵 $\mathbf{D}_{n,m}$，并且由这样确定的矩阵所形成的向量 (3-111) 将被发现是满足 (3-110) 的。

矩阵 $\mathbf{D}_{n,m}$ 显然取代了 § 1.4-3 及 2.3-2 中所考虑的纯量系数 $d_{n,m}$。还可证明当 $h\to 0$ 时 $\mathbf{D}_{n,m}$ 的元素逼近某一个微分方程组的解。令 $\mathbf{d}_{n,m}$ 表示矩阵 $\mathbf{D}_{n,m}$ 的第 k 列 ($k = 1, 2,\cdots,s$)，且令 $\mathbf{e}_k$ 表示第 k 个分量为 1 而其它分量为 0 的向量，于是从 (3-113) 推出

$$\begin{aligned} &\mathbf{d}_{m,m} = \mathbf{e}_k, \\ &\mathbf{d}_{n+1,m} = \mathbf{d}_{n,m} + h\mathbf{G}(x_n)\mathbf{d}_{n,m}. \end{aligned} \tag{3-114}$$

如果用 Euler 方法近似地解初值问题：

$$\begin{aligned} &\mathbf{d}_m(x_m) = \mathbf{e}_k, \\ &\mathbf{d}'_m(x) = \mathbf{G}(x)\mathbf{d}_m(x), \quad x > x_m, \end{aligned} \tag{3-115}$$

所得到的正是同一个方程组。因为 Euler 方法收敛且阶为 1，因此推出

$$\mathbf{d}_{n,m} = \mathbf{d}_m(x_n) + O(h).$$

把对应于 s 个不同的 k 值的 s 个向量 $\mathbf{d}_m(x)$ 组合成矩阵

$\mathbf{D}_m(x)$，我们求得

$$\mathbf{D}_{n,m} = \mathbf{D}_m(x_n) + \boldsymbol{\theta}_{m,n}Kh, \qquad (3\text{-}116)$$

其中 $\boldsymbol{\theta}_{m,n}$ 是一个矩阵，其元素以 1 为界，而 K 为常数，并且矩阵 $\mathbf{D}_m(x)$ 均为初值问题：

$$\begin{aligned}&\mathbf{D}_m(x_m) = \mathbf{I},\\&\mathbf{D}'_m(x) = \mathbf{G}(x)\mathbf{D}_m(x)\end{aligned} \qquad (3\text{-}117)$$

的解．把这个结果应用于一个简单的特殊矩阵 $\mathbf{G}(x)$，例如

$$\mathbf{G}(x) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

我们断定，对于 $s > 1$，矩阵 $\mathbf{D}_{n,m}$ 的元素即使对任意小的 h 值也不再为正．

3.4-3. *累积舍入误差对局部误差的依赖关系；一般情形*．为了用任意一个方法对任意微分方程的近似解导出类似于 (3-111) 的公式，我们将假设局部舍入误差界 ε 按以下方式依赖于 h：

$$Nh^{p+1} \leqslant \varepsilon \leqslant Kh^{q+1}, \quad h \leqslant h_0, \qquad (3\text{-}118)$$

这里 p 表示由 $\boldsymbol{\Phi}$ 所确定的方法的阶，N 为出现在局部离散误差界(3-57)中的常数，K 是不依赖于 h 的常数，并假设 q 满足 $q \geqslant 1$．条件 (3-118) 是可以解释的，例如说，局部舍入误差允许超过局部离散误差，但当 $h \to 0$ 时仍必须至少与 h^2 同样迅速地趋向 0．作为 (3-118) 的一个推理，由定理 3.3 和 3.5，我们有不等式：

$$\begin{aligned}&\|\tilde{\mathbf{y}}_n - \mathbf{y}_n\| \leqslant h^q KE,\\&\|\tilde{\mathbf{y}}_n - \mathbf{y}(x_n)\| \leqslant h^p NE, \quad x_n \in [a, b],\end{aligned} \qquad (3\text{-}119)$$

其中 $E = E_L(b-a)$．

现在我们试图把 (3-107) 写成类似于 (3-110) 的形式．重复使用 Taylor 定理（对于一个及多个变元函数），利用 (3-119) 的第一个关系式，并用 K_1 表示由 (3-90) 给出的 $\boldsymbol{\Phi}$ 的二

阶导数的界，我们有

$$
\begin{aligned}
&\boldsymbol{\Phi}(x_n,\tilde{\mathbf{y}}_n;h)-\boldsymbol{\Phi}(x_n,\mathbf{y}(x_n);h)\\
&\quad=\boldsymbol{\Phi}_j(x_n,\mathbf{y}(x_n);h)(\tilde{y}_n^j-y^j(x_n))\\
&\qquad+\frac{1}{2}\boldsymbol{\Phi}_{jk}(x_n,\mathbf{y}^+;h)((\tilde{y}_n^j-y^j(x_n))(\tilde{y}_n^k-y^k(x_n))\\
&\quad=\mathbf{G}(x_n)(\tilde{\mathbf{y}}_n-\mathbf{y}(x_n))+\varepsilon\boldsymbol{\theta}_nK_2h^q,
\end{aligned}
$$

这里 $\mathbf{y}^+$ 表示对 Φ_{jk} 的不同分量中取不同的中间值的某一个向量．矩阵 $\mathbf{G}(x)$ 如 §3.3-6 中所定义，并且

$$K_2=K_1E(1+2h_0^{q-1}KE).$$

类似地，我们求得

$$
\begin{aligned}
&\boldsymbol{\Phi}(x_n,\mathbf{y}_n;h)-\boldsymbol{\Phi}(x_n,\mathbf{y}(x_n);h)\\
&\qquad=\mathbf{G}(x_n)(\mathbf{y}_n-\mathbf{y}(x_n))+\varepsilon\boldsymbol{\theta}_n'K_3
\end{aligned}
$$

其中

$$K_3=K_1E\left(1+\frac{1}{2}h_0^{q-1}KE\right).$$

于是我们有

$$\boldsymbol{\Phi}(x_n,\tilde{\mathbf{y}}_n;h)-\boldsymbol{\Phi}(x_n,\mathbf{y}_n;h)=\mathbf{G}(x_n)\mathbf{r}_n+\varepsilon\boldsymbol{\theta}_n''K_4, \tag{3-120}$$

其中 $K_4=K_2+K_3$，$\|\boldsymbol{\theta}_n''\|\leqslant 1$． 从而把 (3-107) 可换成为

$$\mathbf{r}_{n+1}=\mathbf{r}_n+h\,\mathbf{G}(x_n)\mathbf{r}_n+\boldsymbol{\epsilon}_{n+1}+\varepsilon\boldsymbol{\theta}_{n+1}hK_4. \tag{3-121}$$

这个关系式等同于 (3-110)，只不过现在以 $\boldsymbol{\epsilon}_{n+1}+\varepsilon\boldsymbol{\theta}_{n+1}h^{q+1}K_4$ 来代替 $\boldsymbol{\epsilon}_{n+1}$． 因此推出

$$\mathbf{r}_n=\mathbf{r}_n^{(1)}+\mathbf{r}_n^{(2)}, \tag{3-122}$$

其中

$$\mathbf{r}_n^{(1)}=\sum_{m=1}^{n}\mathbf{D}_{n,m}\boldsymbol{\epsilon}_m, \tag{3-122a}$$

$$\mathbf{r}_n^{(2)}=\varepsilon\,hK_4\sum_{m=1}^{n}\mathbf{D}_{n,m}\boldsymbol{\theta}_m. \tag{3-122b}$$

和§2.3-2 中一样，我们称 $\mathbf{r}_n^{(1)}$ 为主要舍入误差，称 $\mathbf{r}_n^{(2)}$ 为次要舍入误差． 如果微分方程是线性的并且可用 Euler 方法积分，则只出现主要误差．次要误差的存在性应归结为方程可能为非线性和积分方法的复杂性．从(3-116)和(3-117)推出，矩阵 $\mathbf{D}_{n,m}$ 的元素界是与 h 无关的． 如果 $\mathbf{D}_{n,m}=(d_{n,m}^{ij})$，那么导出

$$\|\mathbf{r}_n^{(2)}\| \leqslant K_5 \varepsilon, \tag{3-123}$$

其中

$$K_5=(b-a)\max_{x_n,x_m\in[a,b]}\left\{\max_{1\leqslant j\leqslant s}\sum_{i=1}^{s}|d_{n,m}^{ij}|\right\},$$

即次要舍入误差当 $h\to 0$ 时与局部舍入误差为同一个数量级．

3.4-4．*累积舍入误差的一个后验界*． 使用以下记号将是方便的． 对于任意向量 $\mathbf{v}=(v^i)$，我们用 $\hat{\mathbf{v}}$ 表示分量为 $|v^i|$ 的向量． 所谓不等式，例如 $\mathbf{v}\leqslant\mathbf{w}$，我们是指对其所有分量有不等式 $v^i\leqslant w^i$ 成立．对于矩阵也使用类似的记号．

在本节中我们将导出 $\hat{\mathbf{r}}_n$ 的界，在假设

$$\hat{\boldsymbol{\varepsilon}}_n\leqslant\mathbf{p}(x_n)\varepsilon \tag{3-124}$$

下，其中 $\mathbf{p}(x)$ 假设为已知向量函数，且具有为 x 的分段连续非负元素．常数 ε 假设满足条件(3-118)． 由于对次要误差已经找到满意的界(3-123)，我们将集中讨论主要误差． 利用(3-122a)，考虑这样的事实，即和单个方程情形一样，$\mathbf{D}_{n,m}$ 的元素不再为正．我们有

$$\hat{\mathbf{r}}_n^{(1)}\leqslant\frac{\varepsilon}{h}\mathbf{m}_n,$$

其中，令 $\mathbf{p}_n=\mathbf{p}(x_n)$，

$$\mathbf{m}_n=h\sum_{m=1}^{n}\hat{\mathbf{D}}_{n,m}\mathbf{p}_m. \tag{3-125}$$

从(3-113)推出

$$\mathbf{D}_{n,n}=\mathbf{I},\mathbf{D}_{n+1,m}\leqslant\mathbf{D}_{n,m}+h\hat{\mathbf{G}}(x_m)\mathbf{D}_{n,m}.$$

采用类似于导出(2-77)的方法,便求得

$$\mathbf{m}_{n+1}-\mathbf{m}_n\leqslant h(\mathbf{p}_{n+1}+\hat{\mathbf{G}}(x_n)\mathbf{m}_n).$$

应用定理 3.3,记住 Euler 方法,我们断定

$$\mathbf{m}_n\leqslant\mathbf{m}(x_n)+O(h), \tag{3-126}$$

其中 $\mathbf{m}(x)$是初值问题

$$\begin{aligned}&\mathbf{m}(a)=0,\\&\dot{\mathbf{m}}(x)=\hat{\mathbf{G}}(x)\mathbf{m}(x)+\mathbf{p}(x)\end{aligned} \tag{3-127}$$

的连续解. 因此我们求得

$$\hat{\mathbf{r}}_n^{(1)}\leqslant\frac{\varepsilon}{h}\{\mathbf{m}(x_n)+O(h)\}. \tag{3-128}$$

并且由于(3-123), 对 $\hat{\mathbf{r}}_n$ 有类似不等式成立. 于是我们证明了:

定理 3.6. 如果增量函数关于它的一切变元其二阶导数连续有界,并且局部舍入误差以(3-124)为界,那么累积舍入误差满足

$$\hat{\mathbf{r}}_n^t\leqslant\frac{\varepsilon}{h}\{\mathbf{m}(x_n)+O(h)\}, \tag{3-129}$$

其中 $\mathbf{m}(x)$ 由(3-127)确定.

对于简单例子 $a=0$, (3-130)

$$y^{1\prime}=y^2,\quad y^{2\prime}=-y^1.$$

假设 $p^i(x)=1$, $i=1,2$, 我们求得关于 $\mathbf{m}(x)$的方程组

$$\begin{aligned}&m^{1\prime}=m^2+1,\quad m^1(0)=0,\\&m^{2\prime}=m^1+1,\quad m^2(0)=0,\end{aligned}$$

其解为

$$m^1(x)=m^2(x)=e^x-1.$$

于是在这种情形,估计式(3-129)允许舍入误差按指数增长,

尽管事实上(3-130)的解保持有界. 这样的情形,对单个线性方程是不会发生的.我们还将看到,这个结果对在下面提出的统计理论是不真实的.

3.4-5. 统计理论. 为了得到关于累积舍入误差的性态的真实结论,我们仍将假设局部舍入误差向量 $\boldsymbol{\epsilon}_n$ 的分量都是随机变量. 关于 $\boldsymbol{\epsilon}_n$ 的随机性将作如下假设:

(i) 令

$$\boldsymbol{\mu}_n = E(\boldsymbol{\epsilon}_n), \tag{3-131}$$

我们假设

$$\hat{\boldsymbol{\mu}}_n \leqslant \mu\, \mathbf{p}(x_n), \tag{3-132}$$

其中 $\mathbf{p}(x)$ 是一个具有非负分量的已知向量,其分量为 x 的分段光滑函数, 并且 $\mu \geqslant 0$ 是与 n 无关的一个常数 (但可能依赖于 h).

(ii) 如果 $m \neq n$, 用上标 T 表示对向量或矩阵的转置, 那么

$$E[(\boldsymbol{\epsilon}_n - \boldsymbol{\mu}_n)(\boldsymbol{\epsilon}_m^T - \boldsymbol{\mu}_m^T)] = 0. \tag{3-133}$$

(iii) 如果 $m = n$, 那么

$$E[(\boldsymbol{\epsilon}_n - \boldsymbol{\mu}_n)(\boldsymbol{\epsilon}_n^T - \boldsymbol{\mu}_n^T)] = \sigma^2\mathbf{C}(x_n), \tag{3-134}$$

其中 $\mathbf{C}(x)$ 是一个已知正半定对称矩阵, 其元素均为 x 的分段光滑函数,并且 σ^2 与 n 无关.

如果把在不同点 x_n 上的局部舍入误差看成独立随机变量,那么条件 (ii) 成立. 在 (3-134) 左端的矩阵称为向量 $\boldsymbol{\epsilon}_n$ 的随机变量的协方差矩阵. 其元素是乘积 $(\epsilon_n^i - \mu_n^i)(\epsilon_n^j - \mu_n^j)$ 的期望值. 如果变量 ϵ_n^i 和 ϵ_n^j 都是独立的, $i \neq j$, 那么协方差矩阵是一个对角矩阵. 在 §3.4-8 中指出,变量 ϵ_n^i 不总是独立的. 因此我们不能假设 $\mathbf{C}(x)$为对角的.

在本节中我们仅涉及到主要误差 $\mathbf{r}_n^{(1)}$,因为难于证明出现在 $\mathbf{r}_n^{(2)}$ 的定义中的向量 $\boldsymbol{\theta}_n$ 的统计假设是正确的.作为 (1-94)

的推理，我们求得 $\mathbf{r}_n^{(1)}$ 的期望值

$$E(\mathbf{r}_n^{(1)}) = \sum_{m=1}^{n} \mathbf{D}_{n,m} \boldsymbol{\mu}_m. \tag{3-135}$$

利用(3-132)，这个向量的分量的绝对值以 $(\mu/h)\mathbf{m}_n$ 的分量为界，其中 $\mathbf{m}_n$ 是由(3-125)确定的。利用前节的结果，于是求得

$$\hat{E}(\mathbf{r}_n^{(1)}) \leqslant \frac{\mu}{h}(\mathbf{m}(x_n) + O(h)), \tag{3-136}$$

其中向量 $\mathbf{m}(x)$ 由(3-127)确定。

现在我们转到确定 $\mathbf{r}_n^{(1)}$ 的协方差矩阵的稍许复杂些的工作。如果

$$E[(\mathbf{r}_n^{(1)} - E(\mathbf{r}_n^{(1)}))(\mathbf{r}_n^{(1)} - E(\mathbf{r}_n^{(1)})^T)] = \mathrm{covar}(\mathbf{r}_n^{(1)}),$$

利用(3-122a)及(3-135)，则有

$$\mathrm{covar}(\mathbf{r}_n^{(1)}) = E\left[\sum_{m=1}^{n} \mathbf{D}_{nm}(\boldsymbol{\varepsilon}_m - \boldsymbol{\mu}_m) \sum_{l=1}^{n} (\boldsymbol{\varepsilon}_l^T - \boldsymbol{\mu}_l^T)\mathbf{D}_{n,l}^T\right]$$

$$= \sum_{m=1}^{n} \sum_{l=1}^{n} \mathbf{D}_{nm} E[(\boldsymbol{\varepsilon}_m - \boldsymbol{\mu}_m)(\boldsymbol{\varepsilon}_l^T - \boldsymbol{\mu}_l^T)]\mathbf{D}_{nl}^T.$$

利用(3-133)及(3-134)，它便化简成

$$\mathrm{covar}(\mathbf{r}_n^{(1)}) = \frac{\sigma^2}{h}\mathbf{V}_n, \tag{3-137}$$

其中，令 $\mathbf{C}_m = \mathbf{C}(x_m)$，

$$\mathbf{r}_n = h\sum_{m=1}^{n} \mathbf{D}_{nm}\mathbf{C}_m\mathbf{D}_{nm}^T. \tag{3-138}$$

象早先一样，我们把 $\mathbf{V}_n$ 通过建立一个差分方程与微分方程的解相比。令 $\mathbf{G}_n = \mathbf{G}(x_n)$，重复利用递推关系式(3-113)，有

$$\mathbf{V}_{n+1} - \mathbf{V}_n = h\left\{\sum_{m=1}^{n+1} \mathbf{D}_{n+1,m}\mathbf{C}_m\mathbf{D}_{n+1,m}^T - \sum_{m=1}^{n} \mathbf{D}_{n,m}\mathbf{C}_m\mathbf{D}_{n,m}^T\right\}$$

$$
= h\left\{\mathbf{C}_{n+1} + \sum_{m=1}^{n} [\mathbf{D}_{n+1,m}\mathbf{C}_m\mathbf{D}_{n+1,m}^T - \mathbf{D}_{nm}\mathbf{C}_m\mathbf{D}_{nm}^T]\right\}
$$

$$
= h\left\{\mathbf{C}_{n+1} + \sum_{m=1}^{n} [(\mathbf{I} + h\mathbf{G}_n)\mathbf{D}_{nm}\mathbf{C}_m\mathbf{D}_{nm}^T(\mathbf{I} + h\mathbf{G}_n^T) - \mathbf{D}_{nm}\mathbf{C}_m\mathbf{D}_{nm}^T]\right\}.
$$

在 $\mathbf{C}(x)$的连续点上，利用以下事实

$$
h^2 \sum_{m=1}^{n} \mathbf{G}_n\mathbf{D}_{nm}\mathbf{C}_m\mathbf{D}_{nm}^T\mathbf{G}_n^T = O(h),
$$

[可从 (3-116) 推出]于是我们求得

$$
\mathbf{V}_{n+1} - \mathbf{V}_n = h\{\mathbf{C}_n + \mathbf{G}_n\mathbf{V}_n + \mathbf{V}_n^T\mathbf{G}_n^T\} + O(h^2). \tag{3-139}
$$

如果用 Euler 方法近似地求解以下矩阵函数 $V(x)$ 的初值问题：

$$
\begin{aligned}
&\mathbf{V}(a) = \mathbf{0},\\
&\mathbf{V}'(x) = \mathbf{C}(x) + \mathbf{G}(x)\mathbf{V}(x) + \mathbf{V}^T(x)\mathbf{G}^T(x),
\end{aligned} \tag{3-140}
$$

则得相同的方程[没有 $O(h^2)$ 项]. 利用定理 3.3, 在 (3-139) 中出现项 $O(h^2)$ 并不影响

$$
\mathbf{V}_n = \mathbf{V}(x_n) + O(h) \tag{3-141}
$$

这一事实的成立，其中 $\mathbf{V}(x)$ 由 (3-140) 所确定. 因此推得

$$
\operatorname{covar}(\mathbf{r}_n^{(1)}) = \frac{\sigma^2}{h}(\mathbf{V}(x_n) + O(h)). \tag{3-142}
$$

我们综合本节结果为以下结论：

定理 3.7. 假设局部舍入误差 $\boldsymbol{\epsilon}_n$ 都是满足在 §3.4-5 中开始叙述的条件 (i), (ii), (iii) 的随机变量，那么累积舍入误差的主要成分 $\mathbf{r}_n^{(1)}$ 的期望值与协方差矩阵满足(3-136) 和 (3-142)，其中 $\mathbf{m}(x)$ 与 $\mathbf{V}(x)$ 由 (3-127) 及 (3-140) 所确定.

在 §3.4-7 中作了关于累积误差分布的论述。

3.4-6. 矩阵 $\mathbf{V}(x)$. (3-140) 的第二个方程表示关于矩阵 $\mathbf{V}(x)$ 的 s^2 个元素的 s^2 个联立微分方程组，因此从 (3-140) 直接计算 $\mathbf{V}(x)$ 是难于实现的. 本节提出的结果可以简化计算 $\mathbf{V}(x)$ 的问题.

引理 3.3. 矩阵 $\mathbf{V}(x)$ 是对称的.

证. 通过对方程组 (3-140) 的转置，因为 $\mathbf{C}^T(x)=\mathbf{C}(x)$，我们求得

$$\mathbf{V}^{T\prime}(x)=\mathbf{C}(x)+\mathbf{V}^T(x)\mathbf{G}^T(x)+\mathbf{G}(x)\mathbf{V}(x).$$

因此 $\mathbf{V}^{T\prime}(x)=\mathbf{V}'(x)$. 由于 $\mathbf{V}^T(0)=\mathbf{V}(0)=\mathbf{0}$，故得结论.

利用上面的结果，积分 s^2 个联立方程便化成积分

$$\frac{1}{2}s(s+1)$$

个方程. 下面给出的结果把问题化成含有 s 个组的积分，而每一个组含有 s 个方程.

引理 3.4. 用 $\mathbf{Y}(x)$ 表示初值问题:

$$\mathbf{Y}(a)=\mathbf{I},\ \mathbf{Y}'(x)=\mathbf{G}(x)\mathbf{Y}(x),\quad a\leqslant x\leqslant b \tag{3-143}$$

的解，那么对于 $a\leqslant x\leqslant b$，$\mathbf{Y}^{-1}(x)$ 是存在的.

注意方程 (3-143) 对于 $\mathbf{Y}(x)$ 的 s 列分成为 s 个组

$$\mathbf{y}'(x)=\mathbf{G}(x)\mathbf{y}(x), \tag{3-144}$$

每一列仅含有 s 个未知函数. 矩阵 $\mathbf{Y}(x)$ 称为组 (3-143) 的完全解.

引理 3.4 的证明. 规定矩阵 $\mathbf{Z}(x)$ 为初值问题

$$\mathbf{Z}(a)=I,\ \mathbf{Z}'(x)=-\mathbf{Z}(x)\mathbf{G}(x),\quad a\leqslant x\leqslant b \tag{3-145}$$

的解，从而

$$(\mathbf{Z}(x)\mathbf{Y}(x))'=\mathbf{Z}'(x)\mathbf{Y}(x)+\mathbf{Z}(x)\mathbf{Y}'(x)$$
$$=-\mathbf{Z}(x)\mathbf{G}(x)\mathbf{Y}(x)+\mathbf{Z}(x)\mathbf{G}(x)\mathbf{Y}(x)=\mathbf{0}.$$

于是

$$\mathbf{Z}(x)\mathbf{Y}(x)=\mathbf{Z}(a)\mathbf{Y}(a)=\mathbf{I},\quad a\leqslant x\leqslant b.$$

这就推得 $\mathbf{Z}(x)$是 $\mathbf{Y}(x)$的左逆矩阵. 因此为 $\mathbf{Y}(x)$的逆矩阵.

引理 3.5[1]. 令 $\mathbf{Y}(x)$由(3-143)所确定,并假设矩阵 $\mathbf{N}(x)$ 满足

$$\mathbf{N}(a)=0,\quad \mathbf{N}'(x)=\mathbf{N}(x)\mathbf{G}^T(x)+\mathbf{Y}^{-1}(x)\mathbf{C}(x),$$
$$a\leqslant x\leqslant b, \tag{3-146}$$

那么

$$\mathbf{V}(x)=\mathbf{Y}(x)\mathbf{N}(x),\quad a\leqslant x\leqslant b. \tag{3-147}$$

注意, (3-146) 的积分可以通过对 $\mathbf{N}(x)$ 的 s 列的每一列具有 s 个未知函数的 s 个方程组的积分来完成.

引理 3.5 的证明. 令 $\mathbf{W}(x)=\mathbf{Y}(x)\mathbf{N}(x)$, 显然

$$\mathbf{W}(a)=\mathbf{V}(a),$$

对于导数,得到

$$\begin{aligned}\mathbf{W}'(x)&=\mathbf{Y}'(x)\mathbf{N}(x)+\mathbf{Y}(x)\mathbf{N}'(x)\\&=\mathbf{G}(x)\mathbf{Y}(x)\mathbf{N}(x)+\mathbf{Y}(x)[\mathbf{N}(x)\mathbf{G}^T(x)\\&\qquad+\mathbf{Y}^{-1}(x)\mathbf{C}(x)]\\&=\mathbf{G}(x)\mathbf{W}(x)+\mathbf{W}(x)\mathbf{G}^T(x)+\mathbf{C}(x).\end{aligned}$$

利用引理 3.3, $\mathbf{V}(x)$满足同一个微分方程. 于是推得

$$\mathbf{W}(x)=\mathbf{V}(x).$$

我们接着来解组 (3-146), 并用完全解来表示.

引理 3.6. (3-146) 的解为

$$\mathbf{N}(x)=\left[\int_a^x \mathbf{Y}^{-1}(t)\mathbf{C}(t)\mathbf{Y}^{-1T}dt\right]\mathbf{Y}^T(x). \tag{3-148}$$

证. 显然有 $\mathbf{N}(a)=0$. 利用对矩阵乘积法则的微分, 我们求得

---

1) 关于这个结果,作者感谢 G. Culler 博士.

$$\mathbf{N}'(x) = \mathbf{Y}^{-1}(x)\mathbf{C}(x)\mathbf{Y}^{-1T}(x)\mathbf{Y}^T(x) + \left[\int_a^x \mathbf{Y}^{-1}(t)\mathbf{C}(t)\mathbf{Y}^{-1T}(t)\,dt\right]\mathbf{Y}^{T\prime}(x) = \mathbf{Y}^{-1}(x)\mathbf{C}(x) + \mathbf{N}(x)\mathbf{G}^T(x).$$

这正是所需要的.

把引理 3.5 与 3.6 的结论结合起来,我们得到以下结果:

定理 3.8. 如果 $\mathbf{Y}(x)$是组 (3-143) 的完全解,由(3-140) 所确定的矩阵 $\mathbf{V}(x)$可表示成如下形式:

$$\mathbf{V}(x) = \mathbf{Y}(x)\left[\int_a^x \mathbf{Y}^{-1}(t)\mathbf{C}(t)\mathbf{Y}^{-1T}(t)\,dt\right]\mathbf{Y}^T(x). \tag{3-149}$$

例. 当

$$\mathbf{G}(x) = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \tag{3-150}$$

时,其中 ω 是一个常数,并设 $a = 0$ 和 $\mathbf{C}(x) = \mathbf{I}$, 我们来确定 $\mathbf{V}(x)$. 易证

$$\mathbf{Y}(x) = \begin{pmatrix} \cos\omega x & \sin\omega x \\ -\sin\omega x & \cos\omega x \end{pmatrix}.$$

因为矩阵 $\mathbf{Y}(x)$是正交的,故有

$$\mathbf{Y}^{-1}(x) = \mathbf{Y}^T(x) \tag{3-151}$$

及

$$\int_a^x \mathbf{Y}^{-1}(t)\mathbf{C}(t)\mathbf{Y}^{-1T}(t)\,dt = x\mathbf{I}.$$

再次利用 (3-151), 我们求得

$$\mathbf{V}(x) = x\mathbf{I}. \tag{3-152}$$

3.4-7. $\mathbf{r}_n^{(1)}$的分布. 我们把前节一些结果应用来讨论当 $n \to \infty$ 时 $\mathbf{r}_n^{(1)}$ 的分量分布. 由 (3-122), 显然 $\mathbf{r}_n^{(1)}$ 的每个分量一般说来是依赖于所有向量 $\boldsymbol{\epsilon}_m$ 的一切分量, $m \leqslant n$. 为了应用定理 1.8, 因此必须假设每一个向量 $\boldsymbol{\epsilon}_n$ 的分量相互之间

都是独立的，或者矩阵 $\mathbf{C}(x_n)$ 是对角线的. 如果我们令

$$\mathbf{V}_n = (v_n^{ij}),\ \mathbf{D}_{nm} = (d_{nm}^{ij}),$$

便发现条件 (1-96) 等价于这样的条件，即对于

$$x_n = x,\ a < x \leqslant b,$$

$$\lim h^{1/2} \frac{d_{nm}^{ij}}{(v_{nm}^{ij})^{1/2}} = 0,\ i, j = 1, \cdots, s \qquad (3-153)$$

对一切 $m < n$ 一致地成立.

关系式 (3-116) 说明，矩阵 $\mathbf{D}_{nm}$ 的元素对于 $a \leqslant x \leqslant b$ 是一致有界的. 从而证明 (3-153) 便变成证明这样的事实，对于一切充分大的 n 值，

$$v_n^{ii} \geqslant c > 0,\ i = 1, \cdots, s, \qquad (3-154)$$

其中 c 与 n 无关. 我们通过证明以下引理来建立 (3-154).

引理 3.7. 如果对称矩阵 $\mathbf{C}(x)$ 对于 $a \leqslant x \leqslant b$ 是分段连续且正定的，那么 $\mathbf{V}(x)$ 对于 $a < x \leqslant b$ 是正定的.

证明[1]是基于重复应用正定矩阵的如下特征.

引理 3.8. 对称矩阵 $\mathbf{A}$ 是正定的当且仅当它可表示成形式 $\mathbf{A} = \mathbf{u}\mathbf{u}^T$, 其中 $\mathbf{u}$ 是非奇异的.

这个证明可参阅 Birkhoff 及 Maclane [1953] 的第 9 章中定理 20.

从引理 3.8 推出，对于 $a \leqslant t \leqslant b$, 我们可求得非奇异矩阵 $\mathbf{B} = \mathbf{B}(t)$, 使得 $\mathbf{C}(t) = \mathbf{B}\mathbf{B}^T$. 两个非奇异矩阵乘积的矩阵 $\mathbf{Y}^{-1}(t)\mathbf{B}(t)$ 是非奇异的. 利用引理 3.8, 矩阵

$$\mathbf{L}(t) = \mathbf{Y}^{-1}(t)\mathbf{C}(t)\mathbf{Y}^{-1T}(t) = \mathbf{Y}^{-1}(t)\mathbf{B}(t)[\mathbf{Y}^{-1}(t)\mathbf{B}(t)]^T$$

是正定的. 因此，如果 $\mathbf{k}$ 是任意非零向量，则有

$$\mathbf{k}^T\mathbf{L}(t)\mathbf{k} > 0,\ a \leqslant t \leqslant b.$$

并且因为 $\mathbf{L}(t)$ 是分段连续的，因此

---

1) 由 G. Cullar 博士所提出的.

$$\int_a^x \mathbf{k}^T \mathbf{L}(t) \mathbf{k}\, dt = \mathbf{k}^T \left[\int_a^x \mathbf{L}(t)\, dt \right] \mathbf{k} > 0, \quad a \leqslant x \leqslant b.$$

于是推得矩阵

$$\mathbf{M}(x) = \int_a^x \mathbf{L}(t)\, dt.$$

对于 $a < x \leqslant b$ 是正定的，并且可分解成形式

$$\mathbf{M}(x) = \mathbf{u}(x)\mathbf{u}^T(x),$$

其中 $\mathbf{u}(x)$ 是非奇异的．再应用引理 3.8，

$$\mathbf{V}(x) = \mathbf{Y}(x)\mathbf{M}(x)\mathbf{Y}^T(x) = \mathbf{Y}(x)\mathbf{u}(x)[\mathbf{Y}(x)\mathbf{u}(x)]^T$$

是正定的．这就证明了引理 3.7.

现在由正定矩阵的对角元素为正这个事实，便可推出关系式 (3-154)．于是证明了 (3-153)，且可陈述为如下结果．

定理 3.9. 假设局部舍入 $\boldsymbol{\epsilon}_n$ 满足在 §3.4-5 开始时所陈述的条件 (i)，(ii) 及 (iii)，其中 $\mathbf{C}(x)$ 是分段连续正定对角矩阵，那么主要舍入误差 $\mathbf{r}_n^{(1)}$ 分量的分布当 $n \to \infty$ 以及 $x_n = x$，$a < x \leqslant b$ 时逼近正态分布．

变步长．以上所建立的容易适用于步长按(2-51)变化的情形，只是我们要求横坐标 x_n 为准确机器数有困难．把步长变因子 $\vartheta(x)$ 对限制为分段常数正函数类．利用类似于 §2.2-8 中给出的分析，可以证明，对于变步长，定理 3.6、3.7 及 3.9 的结论仍然正确，如果向量 $\mathbf{m}(x)$ 与矩阵 $\mathbf{V}(x)$ 的定义 (3-127) 及(3-140) 分别换成

$$\begin{aligned} &\mathbf{m}(a) = \mathbf{0}, \\ &\mathbf{m}'(x) = \mathbf{G}(x)\mathbf{m}(x) + [\vartheta(x)]^{-1}\mathbf{p}(x) \end{aligned} \tag{3-155}$$

及

$$\begin{aligned} \mathbf{V}(a) &= 0, \\ \mathbf{V}'(x) &= [\vartheta(x)]^{-1}\mathbf{C}(x) + \mathbf{G}(x)\mathbf{V}(x) \\ &\quad + \mathbf{V}(x)\mathbf{G}^T(x). \end{aligned} \tag{3-156}$$

3.4-8. 局部舍入误差. 局部舍入误差由(3-104)所确定,它实质上依赖于算术运算是怎样完成的.

(i) 单倍位精确度,定点. 已经给出

$$\boldsymbol{\epsilon}_{n+1}=[h\tilde{\boldsymbol{\Phi}}(x_n,\tilde{\mathbf{y}}_n;h)]^*-h\boldsymbol{\Phi}(x_n,\tilde{\mathbf{y}}_n;h),$$

其中 $\tilde{\boldsymbol{\Phi}}$ 表示 $\boldsymbol{\Phi}$ 的数值近似,以及对于任意向量 $\mathbf{z}$, $\mathbf{z}^*$ 表示一个向量,其分量是对应于 $\mathbf{z}$ 的分量舍入后的值,这正与以前的情形一样,

$$\boldsymbol{\epsilon}_{n+1}=\boldsymbol{\pi}_{n+1}+\boldsymbol{\rho}_{n+1},$$

其中

$$\boldsymbol{\pi}_{n+1}=[h\tilde{\boldsymbol{\Phi}}(x_n,\tilde{\mathbf{y}}_n;h)]^*-h\tilde{\boldsymbol{\Phi}}(x_n,\tilde{\mathbf{y}}_n;h)$$

是引入误差(由乘积 $h\boldsymbol{\Phi}$ 的舍入引起),并且

$$\boldsymbol{\rho}_{n+1}=h[\tilde{\boldsymbol{\Phi}}(x_n,\tilde{\mathbf{y}}_n;h)-\boldsymbol{\Phi}(x_n,\tilde{\mathbf{y}}_n;h)]$$

是固有误差(计算 $\boldsymbol{\Phi}$ 时所固有的). 如果把基本单位记成 u, 那么 $\boldsymbol{\pi}_n$ 的每一个分量以 $\frac{1}{2}u$ 为界, 而 $\boldsymbol{\rho}_{n+1}$ 分量的阶为 hu. 因此, 对于充分小的 h, 局部舍入误差大多数归结为引入误差. 如果将 $\boldsymbol{\pi}_{n+1}$ 的分量当成具有矩形分布 $F_u(x)$ 的随机量[见(1-109)],并且 $\tilde{\boldsymbol{\Phi}}-\boldsymbol{\Phi}$ 被认为是一个随机变量,使得

$$E(\tilde{\boldsymbol{\Phi}}-\boldsymbol{\Phi})=\mathbf{m}u,\quad \mathrm{covar}(\tilde{\boldsymbol{\Phi}}-\boldsymbol{\Phi})=\mathbf{D}u^2,$$

如果 $\mathbf{D}$ 是具有非负元素的对角矩阵, 并且如果变量 $\boldsymbol{\pi}_{n+1}$ 和 $\tilde{\boldsymbol{\Phi}}-\boldsymbol{\Phi}$ 都是独立的,则变量 $\boldsymbol{\epsilon}_{n+1}$ 满足

$$E(\boldsymbol{\epsilon}_{n+1})=\mathbf{m}uh,\quad \mathrm{covar}(\boldsymbol{\epsilon}_{n+1})=\left(\frac{1}{n}\mathbf{I}+h^2\mathbf{D}\right)u^2. \tag{3-157}$$

(ii) 双倍位精确度,定点. 从理论观点来看,用定点十进制或二进制双倍位精确度运算,等价于使用基本单位为 u^2 的单倍位精确度运算.

(iii) 部分双倍位精确度, 定点. 如前,假设 h 和 x_n 都是

精确的单倍位精确度的。用 $\mathbf{z}^*$ 表示对向量 $\mathbf{z}$ 正确舍入后的单倍位精确度的近似值。部分双倍位精确度的算法是由公式：

$$\begin{aligned}&\tilde{\mathbf{y}}_0 = \mathbf{y}_0,\\&\tilde{\mathbf{y}}_{n+1} = \tilde{\mathbf{y}}_n + h\tilde{\boldsymbol{\Phi}}(x_n, \tilde{\mathbf{y}}_n^*; h), \quad n = 0,1,2,\cdots\end{aligned} \tag{3-158}$$

来描述的。

把 $\tilde{\boldsymbol{\Phi}}$ 理解为对 $\boldsymbol{\Phi}$ 的数字单倍位精确度近似(但无需是 $\boldsymbol{\Phi}$ 的正确舍入后的值)。乘积 $h\tilde{\boldsymbol{\Phi}}$ 未经舍入。由于 h 和 $\tilde{\boldsymbol{\Phi}}$ 都表示成精确单倍位精确度的数，从而 $h\tilde{\boldsymbol{\Phi}}$ 的分量和 $\tilde{\mathbf{y}}_n$ 的分量均是精确双倍位精确度的数。在这种情形，局部舍入误差为

$$\boldsymbol{\varepsilon}_{n+1} = h\{\tilde{\boldsymbol{\Phi}}(x_n, \tilde{\mathbf{y}}_n^*; h) - \boldsymbol{\Phi}(x_n, \tilde{\mathbf{y}}_n; h)\}. \tag{3-159}$$

我们有

$$\begin{aligned}\tilde{\boldsymbol{\Phi}}(x_n, \tilde{\mathbf{y}}_n^*; h) - \boldsymbol{\Phi}(x_n, \tilde{\mathbf{y}}_n; h) &= \tilde{\boldsymbol{\Phi}}(x_n, \tilde{\mathbf{y}}_n^*; h) - \boldsymbol{\Phi}(x_n, \tilde{\mathbf{y}}_n^*; h)\\&\quad + \boldsymbol{\Phi}(x_n, \tilde{\mathbf{y}}_n^*; h) - \boldsymbol{\Phi}(x_n, \tilde{\mathbf{y}}_n; h).\end{aligned}$$

用 u 表示基本单位，我们将假设

$$\|\tilde{\boldsymbol{\Phi}} - \boldsymbol{\Phi}\| \leqslant ku,$$

其中 k 是一个固定常数。利用 Lipschitz 条件，(3-56) 和以下事实

$$\|\tilde{\mathbf{y}}_n^* - \mathbf{y}_n\| \leqslant \frac{1}{2}su$$

得到

$$\|\boldsymbol{\varepsilon}_{n+1}\| \leqslant \left(k + \frac{1}{2}sL\right)hu.$$

如果 $u = O(h)$，则推出 (3-118) 成立(因此把 $\mathbf{r}_n$ 分解成主要及次要成分是可能的)。[对于单倍位精确度运算一定要 $u = O(h^2)$。]如果满足这个条件，累积误差性态实质上是通过 (3-122a) 确定的主要误差性态，于是就开辟了一条应用统计理论的途径。

我们假设向量

$$\boldsymbol{\delta}_n = \tilde{\mathbf{y}}_n^* - \tilde{\mathbf{y}}_n \tag{3-160}$$

的分量都是具有分布函数为 $F_u(x)$的独立随机变量，并且向量

$$\boldsymbol{\eta}_n = \tilde{\boldsymbol{\Phi}}(x_n, \tilde{\mathbf{y}}_n^*; h) - \boldsymbol{\Phi}(x_n, \tilde{\mathbf{y}}_n^*; h)$$

的分量都是随机变量，且满足

$$\hat{E}(\boldsymbol{\eta}_n) \leqslant \mathbf{k}\, u, \quad \text{covar}(\boldsymbol{\eta}_n) = \mathbf{W} u^2.$$

向量 $\boldsymbol{\kappa}$ 和矩阵 $\mathbf{W}$ 依赖于 x. 理想的情形是 $\tilde{\boldsymbol{\Phi}}$ 的误差只由舍入而引起，$\mathbf{W} = \frac{1}{12}\mathbf{I}$. 为了求得 $\hat{E}(\boldsymbol{\epsilon}_{n+1})$的界以及 $\text{covar}(\boldsymbol{\epsilon}_{n+1})$ 的表达式，我们注意到利用从 (3-107) 导出 (3-121) 的推理，便有

$$\boldsymbol{\Phi}(x_n, \tilde{\mathbf{y}}_n^*; h) - \boldsymbol{\Phi}(x_n, \tilde{\mathbf{y}}_n; h) = \mathbf{G}(x_n)\delta_n + O(h^2),$$

其中 $\mathbf{G}$ 表示由 (3-120) 确定的函数矩阵. 于是我们有

$$\boldsymbol{\epsilon}_{n+1} = h\{\boldsymbol{\eta}_n + \mathbf{G}(x_n)\boldsymbol{\delta}_n + O(h^2)\}.$$

由于 $E(\boldsymbol{\delta}_n) = 0$，我们求得

$$\hat{E}(\boldsymbol{\epsilon}_{n+1}) \leqslant hu\mathbf{k} + O(h^3). \tag{3-161}$$

此外，由于

$$\begin{aligned}\text{covar}(\mathbf{G}(x_n)\boldsymbol{\delta}_n) &= E(\mathbf{G}(x_n)\boldsymbol{\delta}_n\boldsymbol{\delta}_n^T\mathbf{G}^T(x_n)) \\ &= \mathbf{G}(x_n)E(\boldsymbol{\delta}_n\boldsymbol{\delta}_n^T)\mathbf{G}^T(x_n) = \frac{1}{12}\mathbf{G}(x_n)\mathbf{G}^T(x_n)u^2,\end{aligned}$$

并且因为变量 $\boldsymbol{\eta}_n$ 和 $\boldsymbol{\delta}_n$ 是独立的，我们有

$$\text{covar}(\boldsymbol{\epsilon}_n) = h^2u^2\left\{\mathbf{W} + \frac{1}{12}\mathbf{G}(x_n)\mathbf{G}^T(x_n) + O(h)\right\}. \tag{3-162}$$

方程 (3-161) 和 (3-162) 所确定的量 $\mathbf{p}(x)$ 及 $\mathbf{C}(x)$ 对完成 §3.4-5 中讨论的统计分析是必要的. (iv) 浮点. 令 x 为任意非零实数，并且 b 是 $\geqslant 2$ 的正整数. 对于任意实数 z，用 $[z]$

表示不超过 x 的最大整数，而用 $\log_b$ 表示以 b 为底的对数，令

$$m=[\log_b|x|]+1,\quad a=b^{-m}x, \tag{3-163}$$

于是我们把 x 表示成形式

$$x=ab^m, \tag{3-164}$$

称它为以 b 为底 x 的浮点表示，整数 m 称为指数，实数 a 称为浮点数 x 的尾数。按定义，尾数的绝对值是属于区间 $[b^{-1},1)$.

如果在计算机上表示出一个浮点的数，整数 m 可以(在大范围内)精确地表示. 另一方面，实数 a 一般说来仅能近似地表示，它常常用固定的精确度的数来近似，其基本单位仍记成 u. a 的定点精确度的近似记成 a^*. 如果选取 a^* 使得

$$|a-a^*|\leqslant\frac{1}{2}u,$$

我们则说浮点数是对称地舍入（可惜有些计算系统无对称舍入)。一个浮点数则称为精确机器数，如果 $a=a^*$.

在大型计算机上，几乎普遍采用浮点运算，因为它们无需预估计算中出现的数的大小，以及把问题"换算"到定点运算的范围. 但是，在浮点运算中，不仅对乘积而且对和都是舍去的。

我们称两个浮点数近于相等，如果它们的指数是恒等的. 两个浮点数 x 与 y 的和称为正则的，如果 $x+y$ 的浮点表示的指数等于 x 与 y 中指数的较大者. 一个简单的概率推论证明了在大多数情形，两个不是近于相等的数的和是正则的. 现在我们来讨论 $r=(x+y)^*-(x+y)$，假设 x 及 y 为两个非近于相等的机器数，并且和 $x+y$ 是正则的. 如果

$$|x|>|y|,\quad x=ab^m,$$

并且舍入是对称的，显然有

$$|r|\leqslant\frac{1}{2}ub^m=\frac{1}{2}ub^{[\log_b|x|]+1}. \tag{3-165}$$

虽然上面的界是精确的，但难于进行分析工作，因为出现最大整函数。由于 $m = 1 + [\log_b|x|] = \log_b|x| - \log_b|a|$，我们有

$$b^{[\log_b|x|]+1} = |x|b^{-\log_b|a|}.$$

于是可写成

$$|r| \leqslant u|x|\theta, \tag{3-166}$$

其中 $\theta = \frac{1}{2}b^{-\log_b|a|}$，从而

$$\frac{1}{2} \leqslant \theta < \frac{1}{2}b. \tag{3-167}$$

实际上，θ 可换成 $\frac{1}{2}b$，或者在计算中用适当的平均值来代替。在特别重要的 $b=2$ 这种情形，θ 位于限 $\frac{1}{2}$ 和 1 之间.

现在我们把上面的结果应用到目前的特殊问题。在浮点运算中，递推向量 $\tilde{\mathbf{y}}_n$ 是由公式

$$\begin{aligned} &\tilde{\mathbf{y}}_0 = \boldsymbol{\eta}, \\ &\tilde{\mathbf{y}}_{n+1} = \{\tilde{\mathbf{y}}_n + [h\tilde{\boldsymbol{\Phi}}(x_n, \tilde{\mathbf{y}}_n; h)]^*\}^* \end{aligned} \tag{3-168}$$

生成的。从而局部舍入误差由

$$\boldsymbol{\varepsilon}_{n+1} = \{\tilde{\mathbf{y}}_n + [h\tilde{\boldsymbol{\Phi}}(x_n, \tilde{\mathbf{y}}_n; h)]^*\}^* - \{\tilde{\mathbf{y}}_n + h\boldsymbol{\Phi}(x_n, \tilde{\mathbf{y}}_n; h)\}$$

给出，并且可分成如下的三个分量:

$$\boldsymbol{\varepsilon}_{n+1} = \boldsymbol{\alpha}_{n+1} + \boldsymbol{\pi}_{n+1} + \boldsymbol{\rho}_{n+1}.$$

这里 $\boldsymbol{\pi}_{n+1}$ 和 $\boldsymbol{\rho}_{n+1}$ 如 (i) 中所定义，而称 $\boldsymbol{\alpha}_{n+1}$ 为主导误差，定义为

$$\begin{aligned} \boldsymbol{\alpha}_{n+1} = &\{\tilde{\mathbf{y}}_{n+1} + h[\tilde{\boldsymbol{\Phi}}(x_n, \tilde{\mathbf{y}}_n; h)]^*\}^* \\ &- \{\tilde{\mathbf{y}}_n + [h\tilde{\boldsymbol{\Phi}}(x_n, \tilde{\mathbf{y}}_n; h)]^*\}. \end{aligned} \tag{3-169}$$

让我们来讨论 $\boldsymbol{\alpha}_{n+1}$ 的典型分量 α_{n+1}^i。假设浮点数 y_n^i 和 $h\Phi^i$ 不是近于相等的，其和是正则的，以及 $|y_n^i| > |h\Phi^i|$（这些假设在这里特别适合，因为 h 甚小）。鉴于引入的和固有误差的阶为 $uh\Phi^i$，因而主导误差的阶为 uy_n^i。因此对局部舍入误差

的重大影响是由主导误差给出的。把固有及引入误差完全略去，从而我们可假设

$$\hat{\boldsymbol{\varepsilon}}_{n+1} \leqslant \boldsymbol{\theta}\, u\, \tilde{\mathbf{y}}_n,$$

其中 $\boldsymbol{\theta}$ 为对角矩阵，其对角元素满足 (3-167)。

对于统计工作，更为自然的假设便是把 ε_{n+1} 的分量 ε_{n+1}^i 看成具有分布函数为 $F_w(x)$ 的独立随机变量，其中

$$w = 2u\,\theta|y_n^i|.$$

于是推得 $E(\varepsilon_{n+1}^i) = 0$ 及

$$\operatorname{var}(\varepsilon_{n+1}^i) = \frac{1}{3}u^2\theta^2(y_n^i)^2.$$

3.4-9. 数值例子. 在 §2.3-7 中讨论的数值试验可用以下方式推广到方程组：求线性组

$$\mathbf{y}' = \mathbf{G}(x)\mathbf{y}$$

的数值解族 $\tilde{\mathbf{y}}_{n,q}(q = 1,2,\cdots,Q)$，对应于 Q 个初值条件

$$\mathbf{y}_{0,q} = \mathbf{y}_{0,0}(1 + q\Delta),\ \ q = 1,\cdots,Q.$$

用高精确度的计算来确定理论上的近似值[1] $\mathbf{y}_{n,q}$，对于任何需要的 n 值，于是得到累积舍入误差的 Q 个样本

$$\mathbf{r}_{n,q} = \tilde{\mathbf{y}}_{n,q} - \mathbf{y}_{n,q},$$

借助于关系式

$$\mathbf{a}_n = E(\mathbf{r}_n)_c = \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{r}_{n,q},$$

$$\operatorname{covar}(\mathbf{r}_n)_c = \frac{1}{Q}\sum_{q=1}^{Q}(\mathbf{r}_{n,q} - \mathbf{a}_n)(\mathbf{r}_{n,q}^T - \mathbf{a}_n^T)$$

$$= \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{r}_{n,q}\mathbf{r}_{n,q}^T - \mathbf{a}_n\mathbf{a}_n^T,$$

便可确定值 $\mathbf{r}_{n,q}$ 的均值及其协方差矩阵的试验值。

---

1) 由于 $\mathbf{y}_{n,q} = (1 + q\Delta)\mathbf{y}_{n,0}$，仅需一个 q 值便可确定理论上的近似值。

我们在下面来记录对两个特殊方程组所进行的这个试验结果(每一个包含两个未知函数)。为了简化记号,令

$$y^1 = y,\ y^2 = z,$$

以及

$$E(\mathbf{r}_n) = \begin{pmatrix} p \\ q \end{pmatrix},$$

$$\operatorname{covar}(\mathbf{r}_n) = \begin{pmatrix} a & b \\ b & c \end{pmatrix},\ \mathbf{V}(x) = \begin{pmatrix} u & v \\ v & w \end{pmatrix}.$$

我们满足于对试验的条件和数值结果的简要描述.

试验 1.

微分方程: $y' = -\pi z,\ z' = \pi y.$

初值条件: $x_0 = 0,\ y_{0,0} = 10^{-1} \times 2^{-12},\ z_{0,0} = 0.$

方法: 改进 Euler 方法 [(3-47),取 $\alpha = 1$],单倍位精确度,定点二进制.

$$h = 2^{-6},$$

$$\Delta = \frac{1}{3} \cdot 2^{-8},$$

$$Q = 100,$$

$$u = 2^{-36}.$$

关于局部舍入误差的假设:在 $\left(-\frac{1}{2}u, \frac{1}{2}u\right)$ 内均匀分布,因而 $\sigma^2 = \frac{1}{12}u^2$, $\mathbf{C} = \mathbf{I}$.

$\mathbf{r}_n$ 的理论上的期望值: $p = q = 0.$

协方差矩阵理论上的值:在所考察的情形,矩阵 $\mathbf{V}(x)$ 在 §3.4—6 中计算出来,其结果是

$$u = w = x,\ v = 0.$$

因此我们希望

$$a = c = \frac{1}{12} 2^6 u^2 x, \qquad b = 0. \tag{3-170}$$

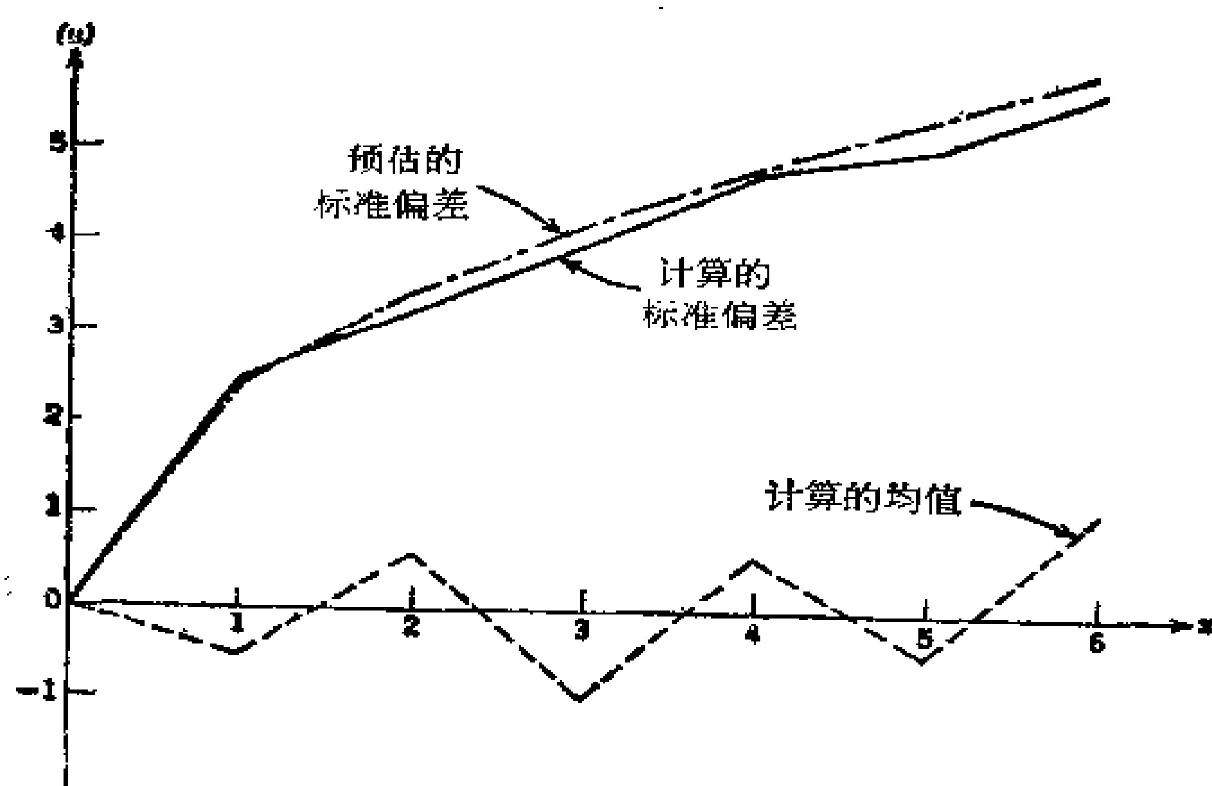


图 3.1　试验 1 中 y_n 值的舍入误差

在表 3.2 中给出 p, q, a, b 及 c 的试验和预估值. 图 3.1 还图解地给出 p 的值及标准偏 $a^{1/2}$ 的值. 预估与试验值之间符合得很好,这是明显的.

表　3.2

	x	1	2	3	4	5	6
试验值	$u^{-1}p$	−0.5	0.6	−0.9	0.6	−0.4	1.1
	$u^{-1}q$	−0.7	0.7	−1.8	2.2	−2.9	3.8
	$u^{-2}a$	5.7	9.7	14.4	20.8	23.2	29.8
	$u^{-2}b$	0.1	−1.3	−2.3	−2.1	−1.5	−4.2
	$u^{-2}c$	5.8	12.7	17.0	21.6	26.9	33.9
预估值	$p=q=b$	0	0	0	0	0	0
	$u^{-2}a=u^{-2}c$	5.3	10.7	16.0	21.3	26.7	32.0

试验 2.

微分方程： $y'=\dfrac{1}{2}z$, $z'=\dfrac{1}{2}y$.

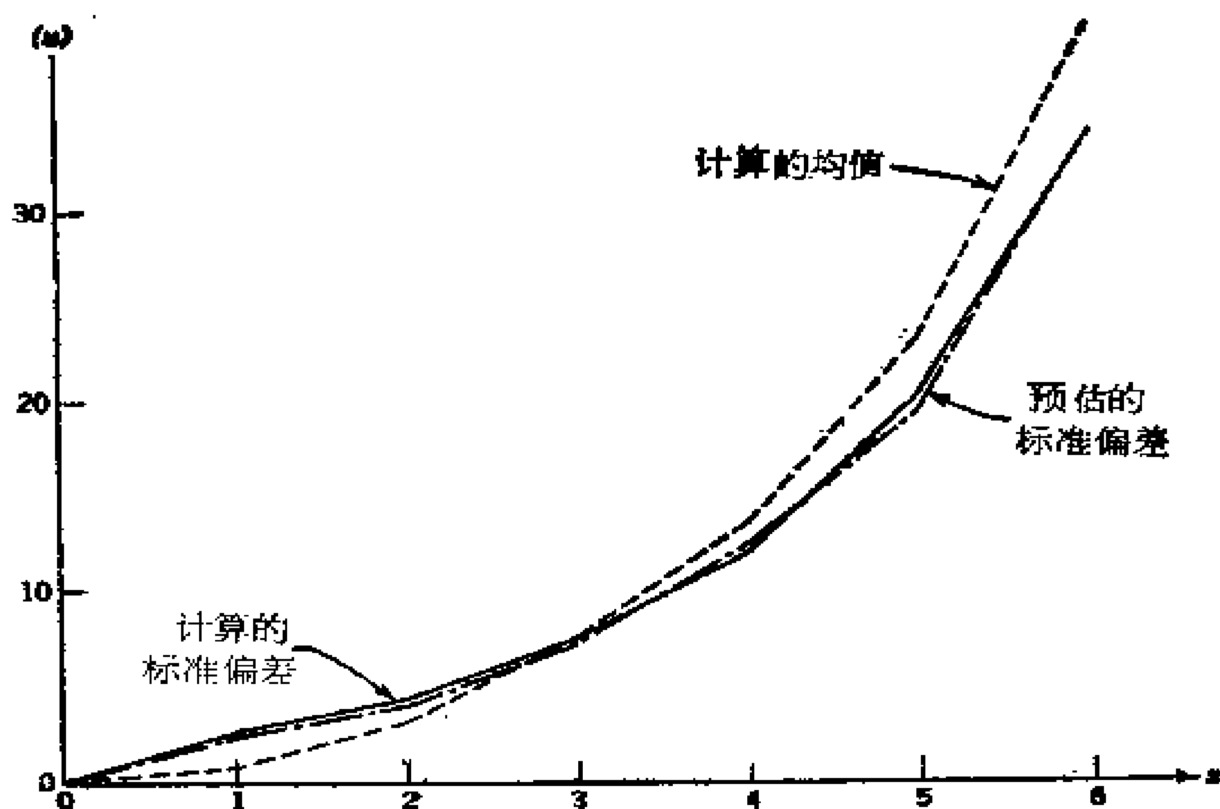


图 3.2

初值条件：$x_0=0$，$y_{0,0}=2^{-17}$，$z_{0,0}=0$。

方法：Runge-Kutta，单倍位精确度，定点二进制．h，Δ，Q及u与试验 1 相同。

关于局部舍入误差的假设：同试验 1。

$\mathbf{r}_n$ 的理论上的期望值：$p=q=0$。

协方差矩阵的理论值：方程（3-140）化成u，v及w的下列方程组：

$$u'=1+v,\ v'=u,\ w'=1+v,$$

其解[在初值条件 $\mathbf{V}(0)=0$ 下]是（见问题 13）

$$u=w=\sinh x,\ v=\cosh x-1. \qquad (3\text{-}171)$$

在表 3.3 中给出 p，q，a，b及 c 的试验值和预估值；在图 3.2 中给出 p 和 $a^{1/2}$ 的值．

显然，这里不能保证零均值的预估是正确的．可以验证

平均误差值与(3-136)所预估的值

$$p=q=\frac{\mu}{n}2(e^{x/2}-1)$$

是十分一致的，如果我们假设 $\mu=uh$. 可以通过假设把对局部舍入误差的分布 $F_{u,v}(x)$ 换成对称分布 $F_u(x)$ 来证明这个

表 3.3

	x	1	2	3	4	5	6
试验值	$u^{-1}p$	0.9	3.2	7.3	13.3	23.0	39.8
	$u^{-1}q$	1.6	3.4	6.8	13.5	22.7	39.0
	$u^{-2}a$	6.8	20.0	57.9	143.1	400.6	1075.1
	$u^{-2}b$	2.3	15.2	51.0	140.8	399.6	1075.2
	$u^{-2}c$	7.0	19.2	53.4	149.7	410.8	1087.3
预估值	$p=q$	0	0	0	0	0	0
	$u^{-2}a=u^{-2}c$	6.3	19.3	53.4	145.6	395.8	1076.3
	$u^{-2}b$	2.9	14.7	58.4	140.3	390.4	1071.0

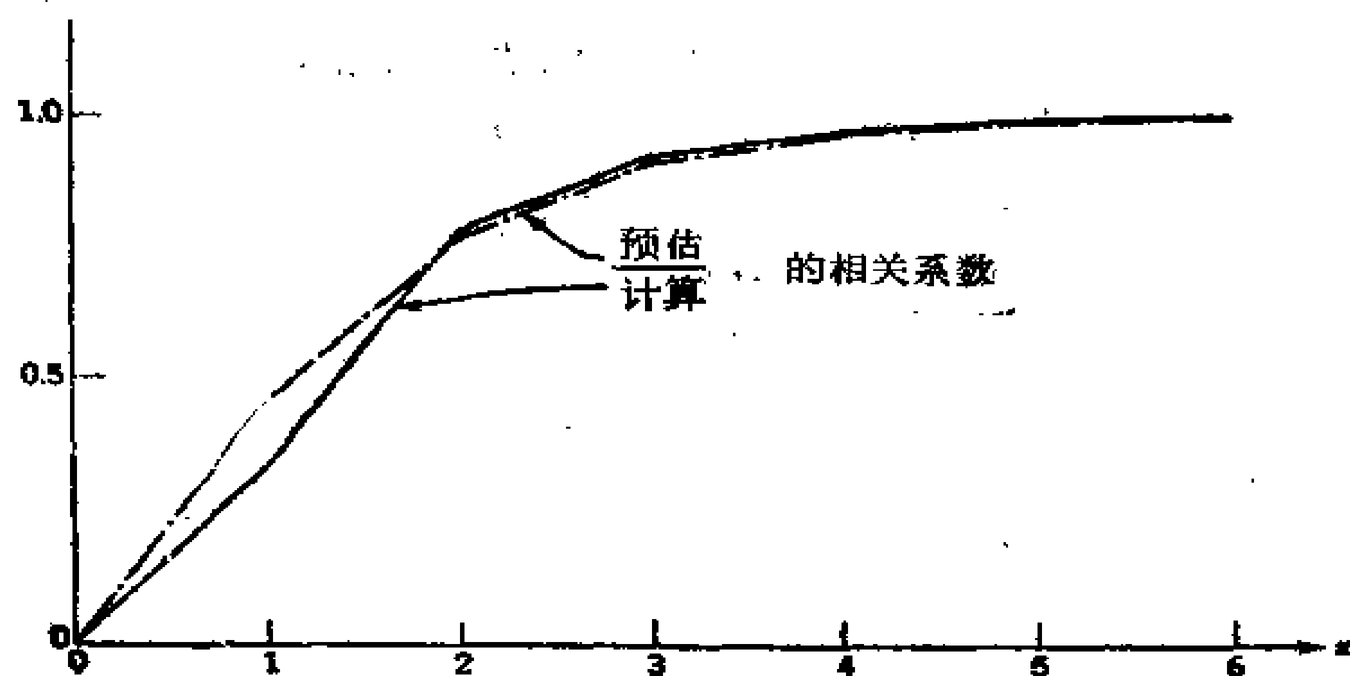


图 3.3 试验 2 中 r_n 的分量之间的相关性

值是正确的，但是由 §1.6-1 预估的 μ 值仅为 $\mu = \frac{1}{2}nh$. 这个偏差的原因还不清楚.

我们从 (3-171) 可以预料到 $\mathbf{r}_n$ 的两个分量是十分相关的. 正如图 3.3 所示，其中相关系数的预估值 $b/(ac)^{1/2}$ 是已知的. 这个预料已由实验准确地证实.

3.5. 求解问题

§3.2.

1. 验证：对于 $s=1$，由 (3-42) 给出的 $\mathbf{f}'(y)$ 和 $\mathbf{f}''(y)$ 的表达式与第二章给出的对应表达式(特别是参看 §2.2-4)相一致.

2. 假设 $s=3$ 并令 $y^1=x, y^2=y, y^3=z; f^1=1, f^2=f, f^3=g$. 全部写出由 (3-42) 确定的量 A^i, B^i, C^i 及 D^i.

3. 令 $y^1=x$, $y^2=y$, $f^1=1$, $f^2=f$, 验证：对于 $s=2$,

$$E^2 = f_{xxx} + 3f_{xxy}f + 3f_{xyy}f^2 + f_{yyy}f^3,$$

$$F^2 = (f_x + f_y f)(f_{yx} + f_{yy}f),$$

$$G^2 = f_y(f_{xx} + 2f_{xy}f + f_{yy}f^2),$$

$$H^2 = f_y^2(f_x + f_y f).$$

4. 确定最一般的三阶 Runge-Kutta 方法，其表达式为

$$\Phi(\mathbf{y};h) = a_1\mathbf{k}_1 + a_2\mathbf{k}_2 + a_3\mathbf{k}_3,$$

其中 $\mathbf{k}_1$, $\mathbf{k}_2$ 和 $\mathbf{k}_3$ 由 (3-48) 所确定. §3.3-7 中所叙述的方法是它的一个特殊情形.

§3.3.

5. 利用张量记号，确定方法的阶与主误差函数，其中 Φ 通过关系式

$$\mathbf{y}_{n+1}=\mathbf{y}_n+\frac{1}{2}h[\mathbf{f}(\mathbf{y}_n)+\mathbf{f}(\mathbf{y}_{n+1})]$$

被隐式地确定。

6*. 对于由

$$\mathbf{y}_{n+1}=\mathbf{y}_n+\frac{1}{2}h[\mathbf{f}(\mathbf{y}_n)+\mathbf{f}(\mathbf{y}_{n+1})]$$
$$+\frac{1}{12}h^2[\mathbf{f}'(\mathbf{y}_n)-\mathbf{f}'(\mathbf{y}_{n+1})]$$

确定的方法，重复问题 5 .

7*. 建立在 Gauss 求积基础上的一个特殊方法． 基于 Gauss 四阶公式的求积方法，可写成

$$\mathbf{y}_{n+1}-\mathbf{y}_n=\frac{1}{2}h[\mathbf{f}(\mathbf{y}_{n+p})+\mathbf{f}(\mathbf{y}_{n+q})], \quad (3\text{-}172)$$

这里 $\mathbf{y}_{n+p}$ 及 $\mathbf{y}_{n+q}$ 分别是在点 x_n+ph 及 x_n+qh 上解的预估值，其中

$$p=\frac{3-\sqrt{3}}{6},\ q=\frac{3+\sqrt{3}}{6}.$$

Hammer 和 Hollingsworth [1955] 提出 $\mathbf{y}_{n+p}$ 及 $\mathbf{y}_{n+q}$ 作为以下方程组的解来确定的：

$$\begin{cases}\mathbf{y}_{n+p}=\mathbf{y}_n+\dfrac{ph}{2(q-p)}[(2q-p)\mathbf{f}(\mathbf{y}_{n+p}) \\ \qquad\qquad -p\,\mathbf{f}(\mathbf{y}_{n+q})], \\ \mathbf{y}_{n+q}=\mathbf{y}_n+\dfrac{qh}{2(q-p)}[q\mathbf{f}(\mathbf{y}_{n+p}) \\ \qquad\qquad -(2p-q)\mathbf{f}(\mathbf{y}_{n+q})].\end{cases} \quad (3\text{-}173)$$

(a) 令

$$\mathbf{y}_{n+p}=\mathbf{y}_n+ph\tilde{\mathbf{\Phi}}_{(1)}(\mathbf{y}_n;ph),$$
$$\mathbf{y}_{n+q}=\mathbf{y}_n+qh\tilde{\mathbf{\Phi}}_{(2)}(\mathbf{y}_n;qh).$$

证明

$$\bar{\Phi}_{(1)}(\mathbf{y};h)=\mathbf{A}+\frac{1}{2}h\mathbf{B}+\frac{1}{6}h^2\gamma_p(\mathbf{C}+\mathbf{D})$$
$$+\frac{1}{24}h^3[\varepsilon_p(\mathbf{E}+3\mathbf{F})+\kappa_p(\mathbf{G}+\mathbf{H})]+O(h^4),$$
$$\bar{\Phi}_{(2)}(\mathbf{y};h)=\mathbf{A}+\frac{1}{2}h\mathbf{B}+\frac{1}{6}h^2\gamma_q(\mathbf{C}+\mathbf{D})$$
$$+\frac{1}{24}h^3[\varepsilon_q(\mathbf{E}+3\mathbf{F})+\kappa_q(\mathbf{G}+\mathbf{H})]+O(h^4),$$
$$(3\text{-}174)$$

其中

$$\gamma_p=\frac{3}{2}\frac{p-a}{p},\ \gamma_q=\frac{3}{2}\frac{q-p}{2},$$
$$\varepsilon_p=2\frac{p^2-pq-q^2}{p^2},\ \varepsilon_q=2\frac{q^2-pq-p^2}{q^2},$$
$$\kappa_p=2\frac{p^2-2pq-q^2}{p^2},\ \kappa_q=2\frac{q^2-2pq-p^2}{q^2}.$$

断定预估公式的阶为 2。

(b) 计算由 (3-172) 及 (3-173) 所确定的方法的增量函数，并证明这个方法的阶为 4 [而不是 3，正如由 (a) 的结果所预期的那样]。

(c) 证明主误差函数为

$$\varphi(\mathbf{y})=-\frac{1}{4320}\mathbf{f}^{\prime\prime}-\frac{1}{108}(\mathbf{K}+\mathbf{L})+\frac{1}{864}(\mathbf{M}+3\mathbf{N})$$
$$-\frac{1}{288}(\mathbf{P}+\mathbf{Q})=-\frac{1}{4320}\mathbf{f}^{\prime\prime}-\frac{1}{864}\mathbf{f}_i f^{\prime\prime\prime i}$$
$$+\frac{1}{432}(\mathbf{f}_{jk}f^j-\mathbf{f}_j f^j_k)f^{\prime\prime k}. \qquad (3\text{-}175)$$

(d) 将这个结果应用于以下方程组[1]的积分：

---

1) 作者感谢 C. B. Tompkins 教授和 D. Pope 博士提供的这个例子，并给出了数值计算。

$$y' = z, \quad y(0) = 0,$$
$$z' = 4y + \frac{1}{10}x^2, \quad z(0) = \frac{1}{10}. \tag{3-176}$$

它的精确解为

$$y = \frac{1}{160}(5e^{2x} - 3e^{-2x} - 4x^2 - 2),$$
$$z = \frac{1}{160}(10e^{2x} + 6e^{-2x} - 8x).$$

尤其是证明了伸缩误差函数的分量 $e = e^1$ 及 $f = e^2$ 为

$$e(x) = -x\left(\frac{1}{720}e^{2x} + \frac{1}{1200}e^{-2x}\right),$$
$$f(x) = -x\left(\frac{1}{360}e^{2x} - \frac{1}{600}e^{-2x}\right).$$

[用网格长 $h = 2^{-p}$, $p = 2,3,\cdots,7$, 在区间 (0,1) 上数值积分,证明在 $x_n = 1$ 处的误差十分准确地按 h^4 法则,有

$$e_n \approx -0.010h^4, \quad f_n = -0.020h^4.$$

与理论一致.]

8. 用在 §3.3-7 中讨论的方法,取步长 $h = 2^{-p}$, $p = 2$, 3,···,对方程组

$$y' = -\pi z, \quad y(0) = 0.1,$$
$$z' = \pi y, \quad z(0) = 0$$

的数值积分得到表 3.4 中给出的数值,并通过伸缩误差而量的估计误差与真正误差相比较.

9. 步长改变因子不以零为界. 证明对于 $a = 0$, $b = 1$, 即使 h 等于 $\frac{1}{2}$, 从 $x = 0$ 至 $x = 1$ 步长改变因子 $\theta = 1 - x$ 要求无限多步数.

§3.4.

10. 证明由 (3-117) 确定的矩阵函数 $\mathbf{D}_m(x)$ 可表示成

表 3.4

	x_n	1	2	3	4	5
y_n	$p=2$	−0.094958995	0.090045421	−0.085265627	0.080624836	−0.076127235
	3	−.099249852	.098504742	−.097764639	.0970229517	−.096299345
	4	−.099902224	.099804542	−.099706952	.099609456	−.099512053
	5	−.099987654	.099975310	−.099962967	.099950626	−.099938286
	6	−.099998453	.099996906	−.099995359	.099993812	−.099992265
	7	−.099999806	.099999613	−.099999419	.099999226	−.099999032
	8	−.099999975	.099999951	−.099999927	.099999903	−.099999879
	9	−.099999996	.099999994	−.099999991	.0999999988	−.099999985
z_n	$p=2$	−0.003559302	0.006759755	−0.009623984	0.012173700	−0.014429704
	3	−.000242890	.000482136	−.000717777	.000949854	−.001178404
	4	−.000015479	.000030929	−.000046349	.000061738	−.000077097
	5	−.000000972	.000001944	−.000002915	.000003887	−.000004858
	6	−.000000061	.000000122	−.000000184	.000000245	−.000000306
	7	−0.000000004	.000000008	−.000000013	.000000017	−.000000021
	8	−.000000000	.000000001	−.000000002	.000000003	−.000000004
	9	−.000000000	.000000001	−.000000001	.000000002	−.000000002

$\mathbf{Y}(x)\mathbf{Y}^{-1}(x_m)$，其中 $\mathbf{Y}(x)$满足

$$\mathbf{Y}'(x) = \mathbf{G}(x)\mathbf{Y}(x), \quad \mathbf{Y}(a) = \mathbf{I}.$$

于是断定

$$\mathbf{D}_{nm} = \mathbf{Y}(x_n)\mathbf{Y}^{-1}(x_m) + O(h).$$

11. 将和 (3-138) 作为近似定积分的 Riemann 和并且使用问题 10 的结果，给出 $V(x)$的表达式的另一个直接证明.

12. 以浮点运算舍入的协方差. 如果 $\mathbf{C}(x)=\mathbf{y}(x)\mathbf{y}^T(x)$，其中 $\mathbf{y}(x)$表示 $\mathbf{y}'(x) = \mathbf{G}(x)\mathbf{y}(x)$的一个解，证明

$$\mathbf{V}(x) = (x - a)\mathbf{y}(x)\mathbf{y}^T(x).$$

作为一个特殊情形，便得到 (3-152).

13. 令 $s = 2$, $\mathbf{C} = \mathbf{I}$ 及

$$\mathbf{G}(x) = \begin{pmatrix} 0 & g(x) \\ g(x) & 0 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} u & v \\ v & w \end{pmatrix},$$

证明

$$u = w = \int_a^x \cosh\left(\int_\xi^x g(t)\,dt\right)d\xi,$$

$$v = \int_a^x \sinh\left(\int_\xi^x g(t)\,dt\right)d\xi.$$

作为一个特例，便得到 (3-171).

14. 令 $s = 2$, $\mathbf{C} = \mathbf{I}$ 及

$$\mathbf{G}(x) = \begin{pmatrix} 0 & g(x) \\ -g(x) & 0 \end{pmatrix},$$

证明

$$\mathbf{V}(x) = (x - a)I.$$

作为一个特殊情形，仍得到 (3-152).

15. 假设局部舍入误差为对称分布并且用定点十进制运算，用单步方法 ($a > 0$)，讨论方程组

$$y' = -z, \quad y(0) = a^{-1},$$

$$z' = -\frac{2}{(x+a)^2}y, \quad z(0) = a^{-2}$$

的解的局部舍入误差传播。

16. 对于方程组

$$y' = -z, \quad y(0) = a^{-1},$$
$$z' = -2y^3, \quad z(0) = a^{-2}$$

重复问题 15.

17. 对于方程组

$$y' = \frac{1}{z}, \quad y(0) = 1,$$

$$z' = -\frac{1}{y}, \quad z(0) = 1,$$

重复问题 15.

18. 假设 $s = 2$。令

$$\mathbf{V} = \begin{pmatrix} u & v \\ v & w \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

其中 a, b, c 及 d 都是常数. 引入向量

$$\mathbf{v} = \begin{pmatrix} u \\ v \\ w \end{pmatrix},$$

方程组 (3-140) 可写成如下形式:

$$\mathbf{v}' = \mathbf{a} + \mathbf{B}\mathbf{v},$$

其中

$$\mathbf{a} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 2a & 2b & 0 \\ c & a+d & b \\ 0 & 2c & 2d \end{pmatrix}.$$

证明对应的齐次方程

$$\mathbf{v}' = \mathbf{B}\mathbf{v}$$

具有形如 $\mathbf{v} = \mathbf{r}e^{(\lambda_i + \lambda_j)x} (i, j = 1, 2)$ 的解，其中 $\mathbf{r}$ 是一个适当

的向量并且 λ_1 和 λ_2 是 $\mathbf{G}$ 的特征值。

19. 令 $\mathbf{G}$ 是一个 s 阶常数矩阵，其特征值为 $\lambda_1,\cdots,\lambda_s$ 而对应的特征向量为 $\mathbf{a}_1,\cdots,\mathbf{a}_s$。 证明对于 s 阶方阵 $\mathbf{y}(x)$ 的微分方程

$$\mathbf{y}' = \mathbf{G}\mathbf{y} + \mathbf{y}\mathbf{G}^T$$

有解 $\mathbf{y} = \mathbf{a}_i\mathbf{a}_j^T e^{(\lambda_i+\lambda_j)x}$, $i, j = 1,\cdots,s$（并不要求这些解都是独立的）。

注

§3.2-3. 本节的分析是模仿 Gill [1951] 的。 一个不同的处理可参看 Albrecht [1955]。 高阶 Runge-Kutta 公式由 Huta [1956,1957] 给出。 关于 Runge-Kutta 方法的其它理论上的贡献，参阅 Kuntzmann [1953,1959a] 及 Ionescu [1954,1956]。计算过程的讨论是由 Murray [1950]，Gill [1951] 及 Blum [1957]， Martin [1958]， Romanelli [1960]， Anderson [1960] 给出的。对 $\mathbf{y}_n$ 的不同分量采用不同步长的 Runge-Kutta 方法的变形是 Rice [1960] 提供的。 Chaplygin 方法的应用（见 §2.1-2 中注）为 Babkin [1954] 及 Artemov [1955] 所讨论。

§3.3-2. 关于离散误差的一般先验界由 Lozinskil [1953] 及Capra [1956] 给出。

§3.3-4. 关于经典 Runge-Kutta 方法对 N 的不同值是由 Bieberbach [1951] 给出的。

§3.3-5. 对于两个方程组用 Heun 方法积分的情形，这里的统计理论是由 Rademacher [1948] 概括的。

第四章　高阶方程组的单步方法

4.1. 引言

4.1-1. q 阶 s 个方程的一般方程组. 本节用 q 表示 $\geqslant 1$ 的一个整数，并且用 $\mathbf{f}, \mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^q$ 表示具有 s 个分量的向量 $(s \geqslant 1)$。我们假设向量值函数 $\mathbf{f}(x, \mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^q)$ 对于 $x \in [a, b]$ 及任意向量 $\mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^q$ 是确定的. q 阶微分方程

$$\mathbf{y}^{(q)} = \mathbf{f}(x, \mathbf{y}', \cdots, \mathbf{y}^{(q-1)}) \tag{4-1}$$

的解是指任意向量值函数 $\mathbf{y}(x)$，它为 q 次可微且满足恒等式

$$\mathbf{y}^{(q)}(x) = \mathbf{f}(x, \mathbf{y}(x), \mathbf{y}'(x), \cdots, \mathbf{y}^{(q-1)}(x)),$$
$$x \in [a, b]. \tag{4-2}$$

在实践中常常会遇到如下的初值问题：求 (4-1) 的解，它满足条件

$$\mathbf{y}(a) = \boldsymbol{\eta}, \mathbf{y}'(a) = \boldsymbol{\eta}', \cdots, \mathbf{y}^{(q-1)}(a) = \boldsymbol{\eta}^{(q-1)}, \tag{4-3}$$

其中 $\boldsymbol{\eta}, \boldsymbol{\eta}', \cdots, \boldsymbol{\eta}^{(q-1)}$ 都是预先指定的向量.

在一定条件下，这个初值问题解的存在唯一性可从定理 3.1 推知. 我们假设

(A) $\mathbf{f}(x, \mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^q)$ 对于 $x \in [a, b]$ 和向量 $\mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^q$ 的分量的任意有限值是确定且连续的；

(B) 存在一个常数 A，使得对任意 $\mathbf{y}^2, \cdots, \mathbf{y}^q$ 和 $\mathbf{y}^{1*}, \cdots, \mathbf{y}^{q*}$ 及一切 $x \in [a, b]$，均有

$$\|\mathbf{f}(x, \mathbf{y}^{1*}, \cdots, \mathbf{y}^{q*}) - \mathbf{f}(x, \mathbf{y}^1, \cdots, \mathbf{y}^q)\|$$
$$\leqslant A(\|\mathbf{y}^{1*} - \mathbf{y}^1\| + \cdots + \|\mathbf{y}^{q*} - \mathbf{y}^q\|). \tag{4-4}$$

即是，$\mathbf{f}$ 关于变元 $\mathbf{y}^1, \cdots, \mathbf{y}^q$ 满足一致的 Lipschitz 条件，于是我们可以证明：

定理 4.1. 令函数 $\mathbf{f}(x, \mathbf{y}^1, \cdots, \mathbf{y}^q)$满足上述条件 (A) 与 (B)，且令向量 $\boldsymbol{\eta}, \boldsymbol{\eta}', \cdots, \boldsymbol{\eta}^{(q-1)}$ 是任意的，那么恰好存在一个函数 $\mathbf{y}(x)$，它在 $[a, b]$上连续且有直到 q 阶的连续导数，并且满足关系式 (4-2) 及 (4-3).

证. 我们规定合成的 sq 维向量为

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}^1 \\ \mathbf{y}^2 \\ \vdots \\ \mathbf{y}^q \end{pmatrix}, \quad \underline{\boldsymbol{\eta}} = \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\eta}' \\ \vdots \\ \boldsymbol{\eta}^{(q-1)} \end{pmatrix},$$

且令 $\mathbf{f}(x, \mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^q) = \mathbf{f}(x, \mathbf{Y})$,

$$\mathbf{f}(x, \mathbf{Y}) = \begin{pmatrix} \mathbf{y}^2 \\ \mathbf{y}^3 \\ \vdots \\ \mathbf{y}^q \\ \mathbf{f}(x, \mathbf{y}) \end{pmatrix}.$$

我们现在考虑含有 sq 个一阶微分方程组的如下初值问题：

$$\mathbf{Y}' = \mathbf{f}(x, \mathbf{Y}), \tag{4-5a}$$

$$\mathbf{Y}(a) = \underline{\boldsymbol{\eta}}. \tag{4-5b}$$

我们把定理 3.1 应用于这个问题，显然满足定理的条件 (A)，并且由

$$\begin{aligned} \|\mathbf{f}(x, \mathbf{Y}^*) - \mathbf{f}(x, \mathbf{Y})\| &\leqslant \|\mathbf{y}^{2*} - \mathbf{y}^2\| + \cdots + \|\mathbf{y}^{q*} - \mathbf{y}^q\| \\ &\quad + A(\|\mathbf{y}^{1*} - \mathbf{y}^1\| + \cdots + \|\mathbf{y}^{q*} - \mathbf{y}^q\|) \\ &\leqslant (1 + A)\|\mathbf{Y}^* - \mathbf{Y}\| \end{aligned}$$

推出满足定理 3.1 的条件 (B)，取 $L = 1 + A$. 因此初值问

题 (4-5) 有唯一解 $\mathbf{Y}(x)$. 如果 $\mathbf{y}^1(x)$ 是表示那个解的第一个子向量，并令 $\mathbf{y}(x)=\mathbf{y}^1(x)$，则由微分方程 (4-5a) 的前 $(q-1)s$ 个分量推出 $\mathbf{y}^{i+1}(x)=\mathbf{y}^{i\prime}(x), i=1,2,\cdots,q-1$. 从而 $\mathbf{y}^{i+1}(x)=\mathbf{y}^{(i)}(x)$, $i=0,1,\cdots,q-1$.

关系式 (4-5b) 包含 (4-3). 考察 (4-5a) 的最后 s 个分量，我们断定 $\mathbf{y}(x)$ 满足 (4-2). 于是这表明了由 (4-1) 及 (4-3) 确定的初值问题解的存在性. 为了证明这个解是唯一的，令 $\mathbf{z}(x)$ 表示初值问题的任意一个解. 容易证明由

$$\mathbf{z}(x)=\begin{pmatrix}\mathbf{z}(x)\\ \mathbf{z}'(x)\\ \vdots\\ \mathbf{z}^{(q-1)}(x)\end{pmatrix}$$

规定的向量 $\mathbf{z}(x)$ 是初值问题 (4-5) 的解. 由于这后一个问题的解的唯一性，$\mathbf{z}(x)$ 恒等于上面所确定的解 $\mathbf{Y}(x)$. 尤其是，前 s 个分量是恒等的. 由此推出 $\mathbf{z}(x)=\mathbf{y}(x)$, $x\in[a,b]$. 这就完成了定理 4.1 的证明.

我们需要关于 $\mathbf{z}(t)$ 的导数 $\mathbf{z}(t), \mathbf{z}'(t), \cdots, \mathbf{z}^{(q-1)}(t)$ 的精确相对增量的一种记法，而 $\mathbf{z}(t)$ 为微分方程

$$\mathbf{z}^{(q)}=\mathbf{f}(t,\mathbf{z},\mathbf{z}',\cdots,\mathbf{z}^{(q-1)}) \tag{4-6a}$$

的解，且满足

$$\mathbf{z}(x)=\mathbf{y}^1, \mathbf{z}'(x)=\mathbf{y}^2, \cdots, \mathbf{z}^{(q-1)}(x)=\mathbf{y}^q, \tag{4-6b}$$

其中 x 是 $[a,b]$ 内固定的点，并且

$$\mathbf{Y}=\begin{pmatrix}\mathbf{y}^1\\ \vdots\\ \mathbf{y}^q\end{pmatrix}$$

是给定的 sq 个向量. 如果

$$\mathbf{z}(t)=\begin{pmatrix}\mathbf{z}(t)\\ \mathbf{z}'(t)\\ \vdots\\ \mathbf{z}^{(q-1)}(t)\end{pmatrix}, \tag{4-7}$$

我们将令

$$\underline{\boldsymbol{\Delta}}(x,\mathbf{Y};h)=\begin{pmatrix}\boldsymbol{\Delta}^1(x,\mathbf{Y};h)\\ \vdots\\ \boldsymbol{\Delta}^q(x,\mathbf{Y};h)\end{pmatrix}$$

$$=\frac{1}{h}[\mathbf{z}(x+h)-\mathbf{z}(x)]. \tag{4-8}$$

4.1-2. 二阶特殊方程组． 在 §4.1-1 中讨论的一般初值问题的如下特殊情形是常常出现的，尤其是天体力学中的问题：

$$\mathbf{y}(a)=\boldsymbol{\eta}, \mathbf{y}'(a)=\eta',\ \mathbf{y}''=\mathbf{f}(x,\mathbf{y}). \tag{4-9}$$

这里的特点是函数并不依赖于一阶导数 $\mathbf{y}'$. 形如 (4-1) 的微分方程被看成特殊的微分方程，其中 $\mathbf{f}$ 不依赖于任何导数.

由于出现各种简化，从数值观点来看，二阶特殊的微分方程也是值得注意的．为了省略标号，我们引入一些记号，它仅限于用在这种特殊方程． 我们始终是令 $\mathbf{y}'=\mathbf{z}$. 引进 $2s$ 个分量向量

$$\mathbf{Y}=\begin{pmatrix}\mathbf{y}\\ \mathbf{z}\end{pmatrix}, \mathbf{f}(x,\mathbf{Y})=\begin{pmatrix}\mathbf{z}\\ \mathbf{f}(x,\mathbf{y})\end{pmatrix}.$$

这 $2s$ 个一阶方程组相当于由 (4-5a) 给出的微分方程 (4-9)。用 $\mathbf{u}(t)$ 表示满足 $\mathbf{u}(x)=\mathbf{y}$, $\mathbf{u}'(x)=\mathbf{z}$ 的 $\mathbf{u}''=\mathbf{f}(t,\mathbf{u})$ 解，其中 $\mathbf{y}$ 和 $\mathbf{z}$ 都是已知向量．我们令

$$\begin{aligned}\mathbf{\Delta}(x,\mathbf{Y};h)&=\frac{\mathbf{u}(x+h)-\mathbf{u}(x)}{h},\\ \theta(x,\mathbf{Y};h)&=\frac{\mathbf{u}'(x+h)-\mathbf{u}'(x)}{h}.\end{aligned}\tag{4-10}$$

因此精确的相对增量的子向量 $\mathbf{\Delta}^1$ 与 $\mathbf{\Delta}^2$ 可记成 $\mathbf{\Delta}$ 与θ.

4.2. 高阶方程组的数值方法

虽然总是可以把高阶方程组化成较大的一阶方程组来进行数值上的处理，但是还可以设计出许多积分方法来直接解决这样的方程组，而无需化成一阶方程组. 基于这些方法中某些具有小的局部离散误差的事实，便要求这些方法也具有小的累积误差. 在 §4.3 中的理论分析将指出这些要求是不合理的. 从离散误差观点来看就没有理由不把高阶方程化成一阶方程组. 由于这个否定的结果，我们仅限于描述，而对问题中某些方法不加证明. 虽然上面关于误差的结论仍然成立，但因为利用这个微分方程取特殊形式的优点便有可能节省一些计算工作，于是对于二阶特殊方程将给予一定的注意.

4.2-1. 一般方程组的增量函数. 方程组 (4-1) 的数值积分的单步方法是由具有 sq 个分量的向量值函数

$$\mathbf{\Phi}(x,\mathbf{Y};h)=\begin{pmatrix}\mathbf{\Phi}^1(x,\mathbf{Y};h)\\ \vdots\\ \mathbf{\Phi}^q(x,\mathbf{Y};h)\end{pmatrix}\tag{4-11}$$

确定的，它使人们利用由公式

$$\begin{aligned}&\mathbf{Y}_0=\boldsymbol{\eta},\\ &\mathbf{Y}_{n+1}=\mathbf{Y}_n+h\mathbf{\Phi}(x_n,\mathbf{Y}_n;h),\ n=0,1,2,\cdots\end{aligned}\tag{4-12}$$

计算出的向量 $\mathbf{y}_n$ 来近似代替 $\mathbf{Y}(x_n)$.

对于每一个函数 $\mathbf{y}^i(x)(i=0,1,\cdots,q-1)$，利用对应

的增量函数的分量 Φ^{i+1} 必须单独地计算出增量. 如果没有得到前 $q-1$ 个导数值,就无法用单步方法来得到解 $\mathbf{y}(x)$的近似. 对于多步方法,情形则不同,见第六章.

对于特殊二阶方程组的数值积分, 我们把增量函数写成形式

$$\boldsymbol{\Phi}(x,\mathbf{Y};h)=\begin{pmatrix}\boldsymbol{\Phi}(x,\mathbf{Y};h)\\ \boldsymbol{\Psi}(x,\mathbf{Y};h)\end{pmatrix}. \tag{4-13}$$

4.2-2. Taylor 展式. 象前几章一样,建立增量函数 $\boldsymbol{\Phi}$ 的一般方法是产生出与所有分量的精确的相对增量 $\boldsymbol{\Delta}$ 尽可能一致. 用显式或隐式 Taylor 展式均能实现这个一致性. 我们仍然用 $\mathbf{f}', \mathbf{f}'', \cdots$ 表示函数 $\mathbf{f}$ 的全导数,这些导数不仅依赖于 $\mathbf{Y}=\mathbf{y}^1$,而且一般地还依赖于 $\mathbf{y}^2,\cdots,\mathbf{y}^q$. 例如,如果微分方程 $\mathbf{y}''=\mathbf{f}(x,\mathbf{y},\mathbf{y}')$ 是已知的,那么以 $y^{\nu i}$ 表示 $\mathbf{y}^\nu(\nu=1,2)$ 的分量,并且采用求和的约定,

$$\mathbf{f}'=\mathbf{f}_x+\mathbf{f}_{y^{1i}}y^{2i}+\mathbf{f}_{y^{2i}}f^i.$$

在特殊二阶方程组 $\mathbf{y}''=\mathbf{f}(x,\mathbf{y})$ 的情形, $\mathbf{f}$ 对 $\mathbf{y}$ 分量的导数可以用 $\mathbf{f}_i$ 明确地表示出来. 并令 $\mathbf{y}^2=\mathbf{z}$, 我们有

$$\mathbf{f}'=\mathbf{f}_x+\mathbf{f}_i z^i.$$

我们将用 $\mathbf{T}^\nu(x,\mathbf{y};h)$ 表示 $\boldsymbol{\Delta}$ 的子向量 $\boldsymbol{\Delta}^\nu$ 的 Taylor 展式部分,而无需用函数 $\mathbf{f}$ 便可计算 $\boldsymbol{\Delta}^\nu$,于是

$$\mathbf{T}^\nu(x,\mathbf{Y};h)=\mathbf{y}^{\nu+1}+\frac{h}{2!}\mathbf{y}^{\nu+2}+\cdots+\frac{h^{q-\nu-1}}{(q-\nu)!}\mathbf{y}^q,$$

$$\nu=1,2,\cdots,q-1,$$

$$\mathbf{T}^q(x,\mathbf{Y};h)=0.$$

从而我们可写成

$$\boldsymbol{\Delta}^\nu(x,\mathbf{Y};h)=\mathbf{T}^\nu(x,\mathbf{Y};h)+\frac{h^{q-\nu}}{(q-\nu+1)!}\mathbf{f}(x,\mathbf{Y})$$
$$+\frac{h^{q-\nu+1}}{(q-\nu+2)!}\mathbf{f}'(x,\mathbf{Y})+\cdots,\nu=1,2,\cdots,q. \tag{4-14}$$

对于特殊二阶方程组，我们有

$$\Delta(x,\mathbf{Y},\mathbf{z};h)=\mathbf{z}+\frac{h}{2}\mathbf{f}(x,\mathbf{Y})+\frac{h^2}{3!}\mathbf{f}'(x,\mathbf{y},\mathbf{z})+\cdots,\tag{4-15}$$

$$\Theta(x,\mathbf{y},\mathbf{z};h)=\mathbf{f}(x,\mathbf{y})+\frac{h}{2}\mathbf{f}'(x,\mathbf{y},\mathbf{z})+\cdots.$$

如果容易计算出一些导数 $\mathbf{f}',\mathbf{f}'',\cdots$，那么在适当的项数以后截断展式(4-14)及(4-15)，便可得到精确的增量函数。但一般来说，并不推荐这个方法.

4.2-3. Runge-Kutta 公式. 把通常 Runge-Kutta 方法(3-48)应用于组 $\mathbf{Y}'=\mathbf{f}(x,\mathbf{Y})$，似乎是浪费，因为仅在 $\mathbf{f}$ 的 sq 个分量的最后 s 个中含有给定的函数 $\mathbf{f}$；而所有其它分量都是不重要的. 因此人们总想计算 Runge-Kutta 系数 $\kappa_\rho(\rho=1,2,\cdots,\kappa)$ 的最后的子向量 κ_ρ 并把增量函数 Φ 的一切子向量 $\Phi^\nu(\nu=1,2,\cdots,q)$ 表示成如下形式：

$$\Phi^\nu(x,\mathbf{Y};h)=\mathbf{T}^\nu(x,\mathbf{Y};h)+\frac{h^{q-\nu}}{(q-\nu+1)!}\sum_{\rho=1}^{\kappa}\gamma_{\nu\rho}\mathbf{k}_\rho,$$

$$\nu=1,2,\cdots,q.\tag{4-16}$$

这里 $\mathbf{k}_\rho$ 表示不依赖于 ν 的函数 $\mathbf{f}$ 的某些值. 选取适当的数值系数 $\gamma_{\nu\rho}$ 使之与 Taylor 展式(4-14)一致（直到一定的幂次).

这个概要的思想是由 Zurmühl [1948] 提出的. 他以形如(4-1)($s=1$)的单个方程的情形得到了 $\mathbf{k}_\rho$ 非常复杂的公式. 由于它们是对称出现的，我们给出如下的稍许不同的公式，它对于形式(4-1)的任意多个方程是成立的. 为了形式上的简化，我们假设 $\mathbf{f}$ 不显式地依赖于 x. 这总是可以办到的，把向量 $\mathbf{y}$ 增加一个分量 y^0，而它满足微分方程

$$y^{0(q)}=0$$

及初值条件

$$y^{(0)}(a)=a,\ y^{0'}(a)=1,\ y^{0(\nu)}(a)=0,\ \nu=2,\cdots,q-1.$$

对于 $q\geqslant 2$，从而推出 $y^{0}(x)=x$. 于是方程组 (4-1) 等价于

$$\mathbf{y}^{(q)}=\mathbf{f}(\mathbf{y},\mathbf{y}',\cdots,\mathbf{y}^{(q-1)}),$$

其中 $\mathbf{y}$ 表示变元向量，并把分量 $f^{0}=0$ 加入到 $\mathbf{f}$.

象 Zurmühl 所作的一样，在 (4-16) 中选取 $\kappa=4$. 我们规定向量 $\mathbf{k}_\rho$ 为

$$\mathbf{k}_\rho=\mathbf{k}_\rho(\mathbf{Y};h)=\mathbf{f}(\mathbf{Y}^1,\mathbf{Y}^2,\cdots,\mathbf{Y}^q),\tag{4-17}$$

其中变元 $\mathbf{Y}^1,\mathbf{Y}^2,\cdots$ 的值由表 4.1 给出.

表 4.1 (4-17) 中 f 的变元

ρ	1	2	3	4
$\mathbf{Y}^1$	$\mathbf{y}^1$	$\mathbf{y}^1+\frac{1}{2}h\,\mathbf{y}^2$	$\mathbf{y}^1+\frac{1}{2}h\,\mathbf{y}^2+\frac{1}{4}h^2\,\mathbf{y}^3$	$\mathbf{y}^1+h\,\mathbf{y}^2+\frac{1}{2}h^2\,\mathbf{y}^3+\frac{1}{4}h^3\,\mathbf{y}^4$
$\mathbf{Y}^2$	$\mathbf{y}^2$	$\mathbf{y}^2+\frac{1}{2}h\,\mathbf{y}^3$	$\mathbf{y}^2+\frac{1}{2}h\,\mathbf{y}^3+\frac{1}{4}h^2\,\mathbf{y}^4$	$\mathbf{y}^2+h\,\mathbf{y}^3+\frac{1}{2}h^2\,\mathbf{y}^4+\frac{1}{4}h^3\,\mathbf{y}^5$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{Y}^{q-3}$	$\mathbf{y}^{q-3}$	$\mathbf{y}^{q-3}+\frac{1}{2}h\mathbf{y}^{q-2}$	$\mathbf{y}^{q-3}+\frac{1}{2}h\mathbf{y}^{q-2}+\frac{1}{4}h^2\mathbf{y}^{q-1}$	$\mathbf{y}^{q-3}+h\,\mathbf{y}^{q-2}+\frac{1}{2}h^2\,\mathbf{y}^{q-1}+\frac{1}{4}h^3\,\mathbf{y}^{q}$
$\mathbf{Y}^{q-2}$	$\mathbf{y}^{q-2}$	$\mathbf{y}^{q-2}+\frac{1}{2}h\mathbf{y}^{q-1}$	$\mathbf{y}^{q-2}+\frac{1}{2}h\,\mathbf{y}^{q-1}+\frac{1}{4}h^2\,\mathbf{y}^{q}$	$\mathbf{y}^{q-2}+h\,\mathbf{y}^{q-1}+\frac{1}{2}h^2\,\mathbf{y}^{q}+\frac{1}{4}h^3\,\mathbf{K}_1$
$\mathbf{Y}^{q-1}$	$\mathbf{y}^{q-1}$	$\mathbf{y}^{q-1}+\frac{1}{2}h\,\mathbf{y}^{q}$	$\mathbf{y}^{q-1}+\frac{1}{2}h\,\mathbf{y}^{q}+\frac{1}{4}h^2\,\mathbf{K}_1$	$\mathbf{y}^{q-1}+h\,\mathbf{y}^{q}+\frac{1}{2}h^2\mathbf{K}_2$
$\mathbf{Y}^{q}$	$\mathbf{y}^{q}$	$\mathbf{y}^{q}+\frac{1}{2}h\,\mathbf{K}_1$	$\mathbf{y}^{q}+\frac{1}{2}h\,\mathbf{K}_2$	$\mathbf{y}^{q}+h\,\mathbf{K}_3$

在 (4-16) 中数值系数 $\varkappa_{\nu\rho}$ 由

$$
\begin{aligned}
\gamma_{\nu 1} &= \frac{(q-\nu+1)^2}{(q-\nu+2)(q-\nu+3)}, \\
\gamma_{\nu 2} = \gamma_{\nu 3} &= \frac{2(q-\nu+1)}{(q-\nu+2)(q-\nu+3)}, \\
\gamma_{\nu 4} &= \frac{1-q+\nu}{(q-\nu+2)(q-\nu+3)}, \quad \nu = 1,2,\cdots,q
\end{aligned} \tag{4-18}
$$

给出. 注意系数 $\gamma_{\nu\rho}$ 仅依赖于 ρ 及 $q-\nu$. 在表 4.2 中给出一些数值.

表 4.2 在 (4-16) 中的系数 $\gamma_{\nu\rho}$ 值

$q-\nu$	$\gamma_{\nu 1}$	$\gamma_{\nu 2}=\gamma_{\nu 3}$	$\gamma_{\nu 4}$
0	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$
1	$\frac{1}{3}$	$\frac{1}{3}$	0
2	$\frac{9}{20}$	$\frac{6}{20}$	$-\frac{1}{20}$
3	$\frac{8}{15}$	$\frac{4}{15}$	$-\frac{1}{15}$

对于非特殊的二阶方程组

$$\mathbf{y}'' = \mathbf{f}(x,\mathbf{y},\mathbf{z}),$$

其中 $\mathbf{z}=\mathbf{y}'$, (4-16) 化成为

$$
\begin{aligned}
\boldsymbol{\Phi}^1(x,\mathbf{y},\mathbf{z};h) &= \mathbf{z} + \frac{1}{6}h(\mathbf{k}_1+\mathbf{k}_2+\mathbf{k}_3), \\
\boldsymbol{\Phi}^2(x,\mathbf{y},\mathbf{z};h) &= \frac{1}{6}(\mathbf{k}_1+2\mathbf{k}_2+2\mathbf{k}_3+\mathbf{k}_4).
\end{aligned} \tag{4-19a}
$$

其中

$$
\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(x,\mathbf{y},\mathbf{z}), \\
\mathbf{k}_2 &= \mathbf{f}\left(x+\frac{1}{2}h,\ \mathbf{y}+\frac{1}{2}h\mathbf{z};\ \mathbf{z}+\frac{1}{2}h\mathbf{k}_1\right), \\
\mathbf{k}_3 &= \mathbf{f}\left(x+\frac{1}{2}h, \mathbf{y}+\frac{1}{2}h\mathbf{z}+\frac{1}{4}h^2\mathbf{k}_1, \mathbf{z}+\frac{1}{2}h\mathbf{k}_2\right), \\
\mathbf{k}_4 &= \mathbf{f}\left(x+h, \mathbf{y}+h\mathbf{z}+\frac{1}{2}h^2\mathbf{k}_2, \mathbf{z}+h\mathbf{k}_3\right).
\end{aligned} \tag{4-19b}
$$

可以证明对于由(4·16)确定的增量函数 $\boldsymbol{\Phi}^{\nu}(\mathbf{x};h)$ 为

$$\boldsymbol{\Phi}^{q}(\mathbf{Y};h)-\boldsymbol{\Delta}^{q}(\mathbf{Y};h)=O(h^{4}), \tag{4-20}$$

$$\boldsymbol{\Phi}^{\nu}(\mathbf{Y};h)-\boldsymbol{\Delta}^{\nu}(\mathbf{Y};h)=O(h^{q-\nu+3}),\nu=1,2,\cdots,q-1.$$

因此，在用相当大的精确度计算出 $\mathbf{y}$ 的值时，$\mathbf{y}^{(q-2)}$ 及 $\mathbf{y}^{(q-1)}$ 的增量函数的精确度是不大于通常的 Runge-Katta 方法的精确度的。

4.2-4. 特殊方程的 Nyström 公式. 对于特殊的二阶方程，$\mathbf{f}$ 的全导数公式稍微简单些，并且使得对给定的置换次数，用 Runge-Kutta 技巧便能达到比一般情形所预期的较高幂次相一致变成为可能. 因此，取两次置换，我们能达到 $\boldsymbol{\Phi}$ 与 $\boldsymbol{\Delta}$ 相一致，误差为 $O(h^3)$；取三次置换，误差为 $O(h^4)$；取四次置换，误差为 $O(h^5)$.

我们仍假设 $\mathbf{f}$ 不显式地依赖于 x，$\mathbf{f}=\mathbf{f}(\mathbf{y})$. 从而增量函数也不依赖于 x，虽然它们除 $\mathbf{y}$ 外是依赖于 $\mathbf{z}=\mathbf{y}'$ 的. 下面讨论的特殊方法是公式

$$\begin{aligned}\boldsymbol{\Phi}(\mathbf{y},\mathbf{z};h)&=\mathbf{z}+h\sum_{\rho=1}^{\kappa}a_{\rho}\mathbf{k}_{\rho},\\ \boldsymbol{\Psi}(\mathbf{y},\mathbf{z};h)&=\sum_{\rho=1}^{\kappa}b_{\rho}\mathbf{k}_{\rho}\end{aligned} \tag{4-21a}$$

的特殊情形，其中

$$\begin{aligned}\mathbf{k}_1&=\mathbf{f}(\mathbf{y}),\\ \mathbf{k}_2&=\mathbf{f}(\mathbf{y}+hp_1\mathbf{z}+h^2p_1q_1\mathbf{k}_1),\\ \mathbf{k}_3&=\mathbf{f}(\mathbf{y}+hp_2\mathbf{z}+h^2p_2[(q_2-q_3)\mathbf{k}_1+q_3\mathbf{k}_2]),\\ \mathbf{k}_4&=\mathbf{f}(\mathbf{y}+hp_3\mathbf{z}+h^2p_3[(q_4-q_5-q_6)\mathbf{k}_1\\ &\qquad+q_5\mathbf{k}_2+q_6\mathbf{k}_3]).\end{aligned} \tag{4-21b}$$

置换数 κ 和参数值 $a_{\rho},b_{\rho}(\rho=1,\cdots,\kappa)$ 以及 $p_1,p_2,\cdots,q_6$ 与所需要的精确度有关：

假设有关函数是充分可微的，如果 p 是最大整数，使得对

一切向量 $\mathbf{y}$，有

$$\Phi(\mathbf{Y};h)-\Delta(\mathbf{Y};h)=O(h^p).$$

我们就称 p 是由 Φ 所确定的方法的阶。可以证明（见问题5），如果 $\kappa=2$，$a_1=a_2=\frac{1}{4}$，$b_1=\frac{1}{4}$，$b_2=\frac{3}{4}$，$p_1=\frac{2}{3}$，$q_1=\frac{1}{2}$ 则 $p=3$。于是公式(4-21)写成

$$\begin{aligned}\Phi(\mathbf{y},\mathbf{z};h)&=\mathbf{z}+\frac{1}{4}h(\mathbf{k}_1+\mathbf{k}_2),\\ \Psi(\mathbf{y},\mathbf{z};h)&=\frac{1}{4}(\mathbf{k}_1+3\mathbf{k}_2),\end{aligned}\tag{4-22a}$$

其中

$$\begin{aligned}\mathbf{k}_1&=\mathbf{f}(\mathbf{y}),\\ \mathbf{k}_2&=\mathbf{f}\left(\mathbf{y}+\frac{2}{3}h\mathbf{z}+\frac{1}{3}h^2\mathbf{k}_1\right).\end{aligned}\tag{4-22b}$$

我们得到 $p=4$（见问题7），如果 $\kappa=3$，且选取参数使得

$$\begin{aligned}\Phi(\mathbf{y},\mathbf{z};h)&=\mathbf{z}+\frac{1}{6}h(\mathbf{k}_1+2\mathbf{k}_2),\\ \Psi(\mathbf{y},\mathbf{z};h)&=\frac{1}{6}(\mathbf{k}_1+4\mathbf{k}_2+\mathbf{k}_3)\end{aligned}\tag{4-23a}$$

其中

$$\begin{aligned}\mathbf{k}_1&=\mathbf{f}(\mathbf{y}),\\ \mathbf{k}_2&=\mathbf{f}\left(\mathbf{y}+\frac{1}{2}h\mathbf{z}+\frac{1}{8}h^2\mathbf{k}_1\right),\\ \mathbf{k}_3&=\mathbf{f}\left(\mathbf{y}+h\mathbf{z}+\frac{1}{2}h^2\mathbf{k}_2\right).\end{aligned}\tag{4-23b}$$

最后有可能达到阶为 $p=5$ 的方法(见问题8)，令

$$\begin{aligned}\Phi(\mathbf{y},\mathbf{z};h)&=\mathbf{z}+\frac{1}{192}h[23\mathbf{k}_1+75\mathbf{k}_2\\ &\qquad-27\mathbf{k}_3+25\mathbf{k}_4],\\ \Psi(\mathbf{y},\mathbf{z};h)&=\frac{1}{192}[23\mathbf{k}_1+125\mathbf{k}_2-81\mathbf{k}_3\\ &\qquad+125\mathbf{k}_4],\end{aligned}\tag{4-24a}$$

其中

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(\mathbf{y}),\\
\mathbf{k}_2 &= \mathbf{f}\left(\mathbf{y} + \frac{2}{5}h\,\mathbf{z} + \frac{2}{25}h^2\mathbf{k}_1\right),\\
\mathbf{k}_3 &= \mathbf{f}\left(\mathbf{y} + \frac{2}{3}h\,\mathbf{z} + \frac{2}{9}h^2\mathbf{k}_1\right),\\
\mathbf{k}_4 &= \mathbf{f}\left(\mathbf{y} + \frac{4}{5}h\,\mathbf{z} + \frac{4}{25}h^2(\mathbf{k}_1 + \mathbf{k}_2)\right).
\end{aligned} \tag{4-24b}$$

公式 (4-22), (4-23) 及 (4-24) 由 Nyström [1925] 在他的关于微分方程数值解的主要研究报告中给出.

4.3. 离散误差

4.3-1. *收敛性，相容性，先验界.* 由增量函数 Φ 所确定方法称为收敛的，如果对于初值向量 η 的每一个的选取，由 (4-12) 确定的向量 $\mathbf{y}_n$ 的分量 $\mathbf{y}_n^1, \mathbf{y}_n^2, \cdots, \mathbf{y}_n^q$ 对 $x \in [a, b]$ 满足

$$\lim_{\substack{h\to 0\\ x_n = x}} \mathbf{y}_n^\nu = \mathbf{y}^{(\nu-1)}(x), \quad \nu = 1, 2, \cdots, q, \tag{4-25}$$

其中 $\mathbf{y}(x)$ 表示初值问题的准确解. 向量 $\mathbf{Y}_n$ 收敛性态的实质可以简单地说成为这些向量起到逼近 sq 个一阶方程组的解的作用. 的确，如果我们总是假设函数 $\Phi(x, \mathbf{Y}; h)$ 在适当的区域内连续，并且关于 $\mathbf{Y}$ 满足 Lipschitz 条件，那么根据定理 3.1 的证明，由 (4-12) 确定的向量 $\mathbf{Y}_n$ 满足

$$\lim_{\substack{h\to 0\\ x_n = x}} \mathbf{y}_n = \mathbf{z}(x), \tag{4-26}$$

其中 $\mathbf{z}(x)$ 表示初值问题

$$\mathbf{z}' = \Phi(x, \mathbf{z}; 0), \quad \mathbf{z}(a) = \eta \tag{4-27}$$

的解.

对于初始向量 η 的一切选取，这个解是与 $\mathbf{y}(x)$ 一致的，

当且仅当

$$\underline{\boldsymbol{\Phi}}(x,\mathbf{z};0)=\mathbf{f}(x,\mathbf{z})$$

对 x 和 $\mathbf{z}$ 为恒等的，即是，如果

$$\begin{aligned}&\boldsymbol{\Phi}^{\nu}(x,\mathbf{Y};0)=\mathbf{y}^{\nu+1},\ \nu=1,2,\cdots q-1,\\&\underline{\boldsymbol{\Phi}}(x,\mathbf{Y};0)=\mathbf{f}(x,\mathbf{Y}).\end{aligned}\tag{4-28}$$

从而这些关系式表达了由 $\underline{\boldsymbol{\Phi}}$ 确定的方法的相容性条件．对于 §4.2 中讨论的特殊方法，它们显然是满足的．对于特殊二阶方程组，相容性条件化成

$$\boldsymbol{\Phi}(x,\mathbf{Y};0)=\mathbf{z},\ \boldsymbol{\Psi}(x,\mathbf{Y};0)=\mathbf{f}(x,\mathbf{y}).\tag{4-29}$$

把定理 3.3 应用于向量 $\mathbf{y}_n$，我们得到误差

$$\mathbf{e}_n=\mathbf{y}_n-\mathbf{y}(x_n)$$

的如下先验界：

定理 4.2. 如果存在一个常数 L，使得，对于任意向量 $\mathbf{y}$ 和 $\mathbf{y}^*$ 以及一切 $x\in[a,b]$ 与 $h\leqslant h_0$，$h_0>0$，均有

$$\|\underline{\boldsymbol{\Phi}}(x,\mathbf{Y}^*;h)-\underline{\boldsymbol{\Phi}}(x,\mathbf{Y};h)\|\leqslant L\|\mathbf{Y}^*-\mathbf{Y}\|,\tag{4-30}$$

并且令 N 和 p 为常数，使得对于 $x\in[a,b]$ 和 $h\leqslant h_0$，有

$$\|\underline{\boldsymbol{\Phi}}(x,\mathbf{Y}(x);h)-\underline{\boldsymbol{\Delta}}(x,\mathbf{Y}(x);h)\|\leqslant Nh^p,\tag{4-31}$$

其中 $\mathbf{y}(x)$ 表示初值问题 (4-5) 的准确解．因此，由 (4-12) 确定的向量 $\mathbf{y}_n$ 满足

$$\begin{aligned}&\|\mathbf{Y}_n-\mathbf{Y}(x_n)\|\leqslant Nh^pE_L(x_n-a),\\&x_n\in[a,b],\ h\leqslant h_0.\end{aligned}\tag{4-32}$$

注意在 (4-31) 中的指数 p 受 $\underline{\boldsymbol{\Phi}}-\underline{\boldsymbol{\Delta}}$ 的分量的近似值的最低次数控制．即是，使得

$$\|\boldsymbol{\Phi}^{\nu}(x,\mathbf{Y}(x);h)-\boldsymbol{\Delta}^{\nu}(x,\mathbf{Y}(x);h)\|\leqslant N_\nu h^{p_\nu},\ h\leqslant h_0\tag{4-33}$$

成立的最小的 p_ν 值的控制．

由于 (4-31) 仅是一个界而不是渐近公式，这个事实本身并不能保证 $\mathbf{y}_n$ 的每一个分量的误差其阶均为 h^p 而不会有任

何更高的阶．但是，在 §4.3-3 中我们将会看到，除极为特殊的例子外，确实是这样的情形．

为了把定理 4.2 的结果应用到特殊方法，则需要知道 L 和 N 的数值．利用 $\mathbf{f}$ 的导数的界，对于这些数值的界可用 §§3.3 和 3.3-4 中的方法求得．对于 Zurmühl 公式，关于 N 的界也可参见 Gautschi [1955]．

4.3-2. 主误差函数．设 $\mathbf{f}$ 和 $\boldsymbol{\Phi}$ 都是充分可微的，我们可把 $\boldsymbol{\Delta}(x,\mathbf{Y};h)$ 与 $\boldsymbol{\Phi}(x,\mathbf{Y};h)$ 按 h 的幂次在任意一点 $(x,\mathbf{y})$，$x\in[a,b]$ 展开．如果 p 是使得对一切 $x\in[a,b]$ 和一切 $\mathbf{y}$ 均有

$$\boldsymbol{\Phi}(x,\mathbf{Y};h)-\boldsymbol{\Delta}(x,\mathbf{Y};h)=O(h^p) \qquad (4\text{-}34)$$

的最大整数，那么称 p 是由 $\boldsymbol{\Phi}$ 所确定的方法的阶．函数

$$\boldsymbol{\varphi}(x,\mathbf{Y})=\begin{pmatrix}\varphi^1(x,\mathbf{Y})\\ \vdots\\ \varphi^q(x,\mathbf{Y})\end{pmatrix} \qquad (4\text{-}35)$$

由

$$\boldsymbol{\Phi}(x,\mathbf{Y};h)-\boldsymbol{\Delta}(x,\mathbf{Y};h)=\boldsymbol{\varphi}(x,\mathbf{Y})h^p+O(h^{p+1}) \quad (4\text{-}36)$$

确定，且称 $\boldsymbol{\varphi}(x,\mathbf{y})$ 为方法的主误差函数．正如前几章一样，主误差函数可用来对局部和累积离散误差渐近地进行估计．

对任何给定的方法，主误差函数作为 $\boldsymbol{\Phi}$ 和 $\boldsymbol{\Delta}$ 的 Taylor 展式之间第一个不同项的偏差是容易计算出来的．我们不采用 §4.2-2 中讨论的方法来实现这个计算．这只要注意到，虽然 $\boldsymbol{\varphi}$ 绝不恒为零，但是 (4-35) 中的一些子向量为零则完全有可能发生．例如，对于方法 (4-16)，我们有

$$\varphi^1=\varphi^2=\cdots=\varphi^{q-2}=0.$$

对于特殊的二阶方程组来说，合适的方法，我们写为

$$\boldsymbol{\varphi}(x,\mathbf{Y})=\begin{pmatrix}\boldsymbol{\varphi}(x,\mathbf{y},\mathbf{z})\\ \boldsymbol{\psi}(x,\mathbf{y},\mathbf{z})\end{pmatrix}.$$

我们将不证明就给出 §4.2-4 中所介绍的一些 Nyström 公式的主误差函数. 我们可假设 $\mathbf{f}$ 不依赖于 x，并利用求和的约定，对于方法 (4-22)，我们有

$$
\begin{aligned}
\boldsymbol{\varphi}(x, \mathbf{z}) &= \frac{1}{72}\mathbf{f}'', \\
\boldsymbol{\Psi}(\mathbf{y}, \mathbf{z}) &= -\frac{1}{648}\mathbf{f}''' - \frac{13}{324}\mathbf{f}_i f^{i\prime},
\end{aligned}
\tag{4-37}
$$

而对于方法 (4-23)，有

$$
\begin{aligned}
\boldsymbol{\varphi}(\mathbf{y}, \mathbf{z}) &= -\frac{1}{720}\mathbf{f}''' - \frac{1}{144}\mathbf{f}_i f^{i\prime}, \\
\boldsymbol{\Psi}(\mathbf{y}, \mathbf{z}) &= \frac{1}{2880}\mathbf{f}^{iv} + \frac{1}{576}\mathbf{f}_i f^{i\prime\prime} + \frac{1}{144}\mathbf{f}_{ij} f^{i\prime} z^j \\
&\quad - \frac{1}{128}\mathbf{f}_{ijk} f^i z^j z^k.
\end{aligned}
\tag{4-38}
$$

对于五阶方法 (4-24)，未计算出主误差函数.

有趣的是可以看到，对于一个简单的特殊情形，例如

$$\mathbf{y}'' = -\mathbf{y},$$

上面的公式是怎样作出的. 在这里我们容易求得

$$\mathbf{f}' = -\mathbf{f}''' = -\mathbf{z} \text{ 及 } \mathbf{f}'' = -\mathbf{f}^{iv} = -\mathbf{y};$$

此外，$\mathbf{f}_i = -\mathbf{e}_i$, $\mathbf{f}_{ij} = \mathbf{f}_{ijk} = \cdots = 0$. 于是对于(4-22)，我们得到

$$\boldsymbol{\varphi} = \frac{1}{72}\mathbf{y}, \quad \boldsymbol{\Psi} = -\frac{1}{24}\mathbf{z}$$

以及对于(4-23)，

$$\boldsymbol{\varphi} = -\frac{1}{120}\mathbf{z}, \boldsymbol{\Psi} = -\frac{1}{480}\mathbf{y}.$$

4.3-3. 累积离散误差的渐近公式. 为了把定理 3.4 应用于方程组 $\mathbf{y}' = f(x, \mathbf{y})$，我们需要知道这个组的函数矩阵 $\mathbf{G}(x)$. 从 $\mathbf{f}(x,\mathbf{y})$ 的定义推出 $\mathbf{G}(x)$ 是 $sq \times sq$ 矩阵

$$
\mathbf{G}(x)=\begin{pmatrix}
0 & \mathbf{I} & & & \\
 & 0 & \mathbf{I} & & \\
 & & \ddots & \ddots & \\
 & & & 0 & \mathbf{I} \\
\mathbf{G}_1(x) & \mathbf{G}_2(x) & \mathbf{G}_3(x) & \cdots & \mathbf{G}_q(x)
\end{pmatrix}, \tag{4-39}
$$

这里 I 表示 $s\times s$ 单位矩阵,并且

$$
\mathbf{G}_\nu=(g_\nu^{ij}(x)),\quad \nu=1,2,\cdots,q,
$$

其中

$$
g_\nu^{ij}(x)=\frac{\partial f^i}{\partial y^{\nu_j}}(x,\mathbf{Y}(x)),\ i,j=1,\cdots,s,
$$

而 y^{ν_j} 表示 $\mathbf{y}^\nu$ 的第 j 个分量. 对于特殊二阶方程组, 由于 $\mathbf{f}$ 不依赖于 $\mathbf{y}^2$, 我们有

$$
\mathbf{G}(x)=\begin{pmatrix}\mathbf{0} & \mathbf{I}\\ \mathbf{G}(x) & \mathbf{0}\end{pmatrix}, \tag{4-39a}
$$

其中 $\mathbf{G}(x)=\mathbf{G}_1(x)$.

于是从定理 3.4 推出

$$
\mathbf{Y}_n-\mathbf{Y}(x_n)=h^p\mathbf{e}(x_n)+O(h^{p+1}), \tag{4-40}
$$

其中 $e(x)$表示初值问题:

$$
\begin{aligned}
\mathbf{e}'(x)=\mathbf{G}(x)\mathbf{e}(x)+\boldsymbol{\varphi}(x,\mathbf{Y}(x)),\\
\mathbf{e}(a)=\mathbf{0}
\end{aligned} \tag{4-41}
$$

的解.

显然,一般说来 $\mathbf{e}(x)$没有恒为零的分量, 即使 $\boldsymbol{\varphi}$ 的一些分量为零. 例如,如果 $\boldsymbol{\varphi}^1=\boldsymbol{\varphi}^2=\cdots=\boldsymbol{\varphi}^{q-1}=\mathbf{0}$, 而 $\boldsymbol{\varphi}^q\not\equiv\mathbf{0}$, 那么,令

$$\mathbf{e}(x)=\begin{pmatrix}\mathbf{e}^1(x)\\ \mathbf{e}^2(x)\\ \vdots\\ \mathbf{e}^q(x)\end{pmatrix}.$$

我们从(4-41)可消去 $\mathbf{e}^2(x),\cdots,\mathbf{e}^q(x)$ 并且规定 $\mathbf{e}^{(1)}(x)$ 为初值问题:

$$\begin{aligned}\mathbf{e}^{1(q)}(x)&=\mathbf{G}_1(x)\mathbf{e}^1(x)+\mathbf{G}_2(x)\mathbf{e}^{1'}(x)+\cdots\\&\quad+\mathbf{G}_q(x)\mathbf{e}^{1(q-1)}(x)+\boldsymbol{\varphi}^q(x,\mathbf{Y}(x)),\\ \mathbf{e}^1(a)&=\mathbf{e}^{1'}(a)=\cdots=\mathbf{e}^{1(q-1)}(a)=\mathbf{0}\end{aligned}\tag{4-42}$$

的解.

于是在这种情形便推出 $\mathbf{y}^{(1)}(x)$ 的误差渐近地为 $O(h^p)$,即使 $\boldsymbol{\Phi}^1-\boldsymbol{\Delta}^1$ 具有更高些的阶. 关于一个特殊微分方程在 Zurmühl 公式的特殊情形下,这个重要的否定结果为 Rutishauser [1955] 所说明.

对于特殊二阶微分方程组更完全的分析是可能的. 如果我们令

$$\mathbf{e}(x)=\begin{pmatrix}\mathbf{e}(x)\\ \mathbf{d}(x)\end{pmatrix},$$

那么方程组(4-41)可写成

$$\begin{aligned}&\mathbf{e}'(x)=\mathbf{d}(x)+\boldsymbol{\varphi}(x,\mathbf{Y}(x)),\\&\mathbf{d}'(x)=\mathbf{G}(x)\mathbf{e}(x)+\boldsymbol{\phi}(x,\mathbf{Y}(x)),\\&\mathbf{e}(a)=\mathbf{d}(a)=\mathbf{0}.\end{aligned}\tag{4-43}$$

消去 $\mathbf{d}(x)$ 便得到

$$\begin{aligned}&\mathbf{e}''(x)=\mathbf{G}(x)\mathbf{e}(x)+\boldsymbol{\phi}(x,\mathbf{Y}(x))+\boldsymbol{\varphi}'(x,\mathbf{Y}(x)),\\&\mathbf{e}(a)=\mathbf{0},\mathbf{e}'(a)=\boldsymbol{\varphi}(a,\mathbf{y}(a)).\end{aligned}\tag{4-44}$$

于是推出,仅在 $\boldsymbol{\phi}(x,\mathbf{Y}(x))+\boldsymbol{\varphi}'(x,\mathbf{Y}(x))=\mathbf{0}$ 及

$$\boldsymbol{\varphi}(a,\mathbf{Y}(a))=\mathbf{0}$$

的特殊情形,$\mathbf{e}(x)=\mathbf{0}$.

上述结果可正式陈述如下：

定理 4.3. 如果由 $\boldsymbol{\Phi}(x, \mathbf{Y}; h)$ 所确定的方法为 p 阶，并且 $\mathbf{f}$ 和 $\boldsymbol{\Phi}$ 对 $x \in [a, b]$ 及任意 $\mathbf{y}$ 都有 $p+2$ 阶连续导数，则初值问题 (4-5) 的解的误差 $\mathbf{e}_n = \mathbf{Y}_n - \mathbf{Y}(x_n)$ 当 $x_n \in [a, b]$，$h \to 0$ 时满足

$$\mathbf{e}_n = h^p \mathbf{e}(x_n) + O(h^{p+1}), \tag{4-45}$$

其中 $\mathbf{e}(x)$ 是初值问题 (4-41) 的解.

为了便于说明，我们把上面的结果应用于通常的初值问题 $y'' = -y$，$y(0) = 1$，$y'(0) = 1$，这里 $s = 1$. 对于特殊的 Nyström 公式 (4-23)，我们有 $p = 4$，并利用 (4-44)，有

$$e''(x) = -e(x) + \frac{1}{160} y(x) = -e(x) + \frac{1}{160} \cos x,$$

$$e(0) = e'(0) = 0.$$

于是

$$e(x) = \frac{1}{320}(x \cos x - \sin x).$$

把通常的 Runge-Kutta 方法 (3-48) 应用于方程组 $y' = z$，$z' = -y$，$y(0) = 1$，$z(0) = 0$，我们便有(用本章的记号)

$\varphi = -\frac{1}{120} z$，$\psi = \frac{1}{120} y$，于是

$$e''(x) = -e(x) + \frac{1}{60} \cos x,$$

$$e(0) = e'(0) = 0.$$

这就导出略差一些的结果

$$e(x) = \frac{1}{120}(x \cos x - \sin x).$$

4.4. 舍入误差传播

4.4-1. 非统计理论. 令数值 $\tilde{\mathbf{Y}}_n$ 满足

$$\tilde{\mathbf{Y}}_{n+1} = \mathbf{Y}_n + h\boldsymbol{\Phi}(x_n, \mathbf{Y}_n; h) + \boldsymbol{\varepsilon}_{n+1}, \tag{4-46}$$

其中向量 $\boldsymbol{\varepsilon}_{n+1}$ 表示局部舍入误差．从定理 3.5 可得累积舍入误差 $\mathbf{r}_n = \tilde{\mathbf{Y}}_n - \mathbf{Y}_n$ 的一个粗糙界：如果局部舍入误差满足

$$\|\boldsymbol{\varepsilon}_n\| \leqslant \varepsilon, \quad n = 1, 2, \cdots \tag{4-47}$$

并且 L 表示由 (4-30) 确定的 Lipschitz 常数，那么我们显然有

$$\|\mathbf{r}_n\| \leqslant \frac{\varepsilon}{h} E_L(x_n - a), \quad x_n \in [a, b]. \tag{4-48}$$

假设

$$\hat{\boldsymbol{\varepsilon}}_n \leqslant \varepsilon \mathbf{P}(x_n), \tag{4-49}$$

便可得到更为精确的界，其中 $\mathbf{P}(x)$ 是固定的具有非负分量的分段连续和连续可微的函数，并且

$$Nh^{p+1} \leqslant \varepsilon \leqslant Kh^{q+1}, \tag{4-50}$$

这里 N, K 和 q 都是不依赖于 h 的常数，$1 \leqslant q \leqslant p$．当允许局部舍入误差支配局部离散误差时，假设 (4-50) 则是很自然的．于是从定理 3.6 及以前的讨论便可推出

$$\mathbf{r}_n = \mathbf{r}_n^{(1)} + \mathbf{r}_n^{(2)},$$

其中 $\mathbf{r}_n^{(2)} = O(\varepsilon)$，并且

$$\hat{\mathbf{r}}_n^{(1)} \leqslant \frac{\varepsilon}{h} \{\mathbf{m}(x_n) + O(h)\}. \tag{4-51}$$

函数 $\mathbf{m}(x)$ 规定为初值问题

$$\mathbf{m}' = \hat{\mathbf{G}}(x)\mathbf{m} + \mathbf{P}(x), \quad \mathbf{m}(a) = 0 \tag{4-52}$$

的解．对特殊二阶方程组的情形，我们令

$$\mathbf{r}_n^{(1)} = \begin{pmatrix} \mathbf{r}_n^{(1)} \\ \mathbf{s}_n^{(1)} \end{pmatrix}, \quad \mathbf{m} = \begin{pmatrix} \mathbf{m} \\ \mathbf{n} \end{pmatrix}, \mathbf{P}(x) = \begin{pmatrix} \mathbf{p}(x) \\ \mathbf{q}(x) \end{pmatrix}.$$

由于矩阵 $\mathbf{G}$ 的特殊形式 (4-39a)，方程组 (4-52) 就具有形式：

$$
\begin{aligned}
&\mathbf{m}' = \mathbf{n} + \mathbf{p}(x), \\
&\mathbf{n}' = \hat{\mathbf{G}}(x)\mathbf{m} + \mathbf{q}(x), \\
&\mathbf{m}(a) = \mathbf{n}(a) = \mathbf{0}.
\end{aligned}
\tag{4-53}
$$

如果仅对 $\mathbf{r}_n^{(1)}$ (即 y_n 的舍入误差)的增长有兴趣，那么可从 (4-53) 中消去 $\mathbf{n}(x)$，并且求得

$$
\hat{\mathbf{r}}_n^{(1)} \leqslant \frac{\varepsilon}{h}\{\mathbf{m}(x_n) + O(h)\}, \tag{4-54}
$$

其中

$$
\begin{aligned}
&\mathbf{m}'' = \hat{\mathbf{G}}(x)\mathbf{m} + \mathbf{p}'(x) + \mathbf{q}(x), \\
&\mathbf{m}(a) = 0, \ \mathbf{m}'(a) = \mathbf{p}(a).
\end{aligned}
\tag{4-55}
$$

4.4-2. 统计理论. §3.4-5 的结果容易应用于目前的情形. 我们把舍入误差 $\boldsymbol{\varepsilon}_n$ 当作随机变量,其期望值为

$$
\boldsymbol{\mu}_n = E(\boldsymbol{\varepsilon}_n).
$$

并假设

$$
\hat{\boldsymbol{\mu}}_n \leqslant \mu \mathbf{P}(x_n),
$$

$$
E[(\boldsymbol{\varepsilon}_n - \boldsymbol{\mu}_n)(\boldsymbol{\varepsilon}_m^T - \boldsymbol{\mu}_m^T)] = \begin{cases} \mathbf{0}, & n \neq m, \\ \sigma^2 \mathbf{C}(x_n), \end{cases} \tag{4-56}
$$

其中 $\mathbf{P}(x)$ 是已知的分段连续且连续可微的向量函数，而 $\mathbf{C}(x)$是已知的正半定矩阵，其元素为具有类似性质的 x 的函数。把定理 3.7 应用于初值问题 (4-5)，于是求得

$$
\hat{E}(\mathbf{r}_n^{(1)}) \leqslant \frac{\mu}{h}\{\mathbf{m}(x_n) + O(h)\}, \tag{4-57}
$$

其中 $\mathbf{m}(x)$ 由 (4-52) 确定. 尤其是,对于特殊的二阶方程组,

$$
\hat{E}(\mathbf{r}_n^{(1)}) \leqslant \frac{\mu}{h}\{\mathbf{m}(x_n) + O(h)\} \tag{4-58}
$$

而 $\mathbf{m}(x)$ 由 (4-55) 所确定.

对于协方差矩阵,其结果更为复杂. 定理 3.7 陈述为

$$
\operatorname{covar}\{\mathbf{r}_n^{(1)}\} = \frac{\sigma^2}{h}\{\mathbf{V}(x_n) + O(h)\}, \tag{4-59}
$$

其中 $\mathbf{V}(x)$ 定义为如下 s^2q 个微分方程组

$$\mathbf{V}' = \mathbf{C} + \mathbf{G}\mathbf{V} + \mathbf{V}\mathbf{G}^T; \quad \mathbf{V}(a) = 0 \tag{4-60}$$

的解. 对于特殊的二阶方程组,我们令

$$\operatorname{covar}(\mathbf{r}_n^{(1)}) = \begin{pmatrix} \mathbf{R} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{T} \end{pmatrix}, \mathbf{C} = \begin{pmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{D}^T & \mathbf{E} \end{pmatrix}, \mathbf{V} = \begin{pmatrix} \mathbf{V} & \mathbf{W} \\ \mathbf{W}^T & \mathbf{U} \end{pmatrix}.$$

于是方程组 (4-60) 可写成形式

$$\mathbf{V}' = \mathbf{C} + \mathbf{W}^T + \mathbf{W}, \tag{4-61}$$

$$\begin{aligned} &\mathbf{W}' = \mathbf{D} + \mathbf{U} + \mathbf{V}\mathbf{G}^T, \\ &\mathbf{U}' = \mathbf{E} + \mathbf{G}\mathbf{W} + \mathbf{W}^T\mathbf{G}^T, \end{aligned} \qquad \mathbf{V}(a) = \mathbf{W}(a) = \mathbf{U}(a) = \mathbf{0}.$$

如果仅对 $\mathbf{r}_n^{(1)}$ 的协方差有兴趣,那么只需要确定矩阵 $\mathbf{V}$. 但是,用任何简单的方式从组中消去 $\mathbf{W}$ 和 $\mathbf{U}$ 几乎是不可能的,除 $s = 1$ 的情形外,其中矩阵 $\mathbf{C}, \mathbf{D}, \cdots$ 化成标量. 如果

$$\mathbf{C} = \begin{pmatrix} c & d \\ d & e \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} v & w \\ w & u \end{pmatrix},$$

方程组 (4-61) 便化成

$$\begin{aligned} v' &= c + 2w, \\ w' &= d + u + gv, \\ u' &= e + 2gw, \end{aligned} \tag{4-62}$$

其中 $g = g(x) = f_y(x, y(x))$. 假设 $g(x)$ 和 $\mathbf{c}$ 的元素都是充分可微的,我们求得

$$\begin{aligned} v'' &= c' + 2w' = c' + 2(d + u + gv), \\ v''' &= c'' + 2(d' + u' + (gv)') \\ &= c'' + 2(d' + e + g(v' - c) + (gv)'). \end{aligned}$$

从而 $v(x)$ 满足单个微分方程

$$v''' - 4gv' - 2g'v = c'' + 2d' + 2e - 2gc \tag{4-63a}$$

以及初值条件

$$\begin{aligned} &v(a) = 0, \quad v'(a) = c(a), \\ &v''(a) = c'(0) + 2d(0). \end{aligned} \tag{4-63b}$$

对于我们在 §4.3-3 中已经讨论过的初值问题 $y''=-y$，$y(0)=1$，$y'(0)=0$，有 $g(x)=-1$。假设 $c=e=1$，$d=0$（常数对角的局部协方差矩阵，适合定点运算），初值问题（4-63）化成

$$v'''+4v'=4,\quad v(0)=v''(0)=0;\quad v'(0)=1.$$

根据（3-152），容易验证其解为 $v(x)=x$。

4.5. 求解的问题

§4.2.

1. 利用 Nyström 公式（4-22），取步长 $h=0.4$ 和 $h=0.2$，求初值问题

$$y''=-\sin y,\quad y(0)=1,\quad y'(0)=0$$

的解 $y(x)$，并且用反插值（见 Hildebrand [1956]，第 50 页）确定 $y(x)$ 的最小正根 ξ。

[对于 $y(0)=\alpha$，由

$$\xi=\frac{\pi}{2}\left[1+\left(\frac{1}{2}\right)^2\left(\sin\frac{\alpha}{2}\right)^2+\left(\frac{1.3}{2.4}\right)^2\left(\sin\frac{\alpha}{2}\right)^4\right]+\cdots$$

给出一个解析解。]

2. 采用 Nyström 公式（4-23）的数值积分，求出由

$$y''+\frac{4x^2}{1+x^2}y=0,\quad y(0)=0,\quad y'(0)=1$$

确定的正弦状曲线的渐近振幅与相位改变。

3. 采用这样方法来确定参数 p_1，使公式（要求一次 f 置换）

$$\boldsymbol{\Phi}(\mathbf{y},\mathbf{z};h)=\mathbf{z}+\frac{1}{2}h\mathbf{k},$$

$$\boldsymbol{\Psi}(\mathbf{y},\mathbf{z};h)=\mathbf{k}$$

表示特殊二阶方程组的二阶方法，其中 $\mathbf{k}=\mathbf{f}(\mathbf{y}+hp_1\mathbf{z})$。

4. 令 $\mathbf{f} = \mathbf{f}(\mathbf{y})$，$\mathbf{f}_i = \partial \mathbf{f}/\partial y^i$，并利用求和约定，

$$\mathbf{A} = \mathbf{f}, \qquad \mathbf{H} = \mathbf{f}_{ijkm}\, z^i z^j z^k z^m,$$
$$\mathbf{B} = \mathbf{f}_i\, z^i, \qquad \mathbf{I} = \mathbf{f}_{ijk}\, f^i z^j z^k,$$
$$\mathbf{C} = \mathbf{f}_{ij}\, z^i z^j, \qquad \mathbf{J} = \mathbf{f}_{ij}\, f^i_k z^j z^k$$
$$\mathbf{D} = \mathbf{f}_i f^i, \qquad \mathbf{K} = \mathbf{f}_{ij}\, f^i f^j,$$
$$\mathbf{E} = \mathbf{f}_{ijk}\, z^i z^j z^k, \qquad \mathbf{L} = f_i f^i_{jk} z^j z^k,$$
$$\mathbf{F} = \mathbf{f}_{ij}\, f^i z^j \qquad \mathbf{M} = f_i f^i_j f^j.$$
$$\mathbf{G} = \mathbf{f}_i\, f^i_j z^j,$$

如果 $\mathbf{y}' = \mathbf{z}$，$\mathbf{y}'' = \mathbf{f}(\mathbf{y})$，证明 $\mathbf{f}$ 对于 x 的全导数为

$$\mathbf{f}' = \mathbf{B},$$
$$\mathbf{f}'' = \mathbf{C} + \mathbf{D},$$
$$\mathbf{f}''' = \mathbf{E} + 3\mathbf{F} + \mathbf{G},$$
$$\mathbf{f}^{iv} = \mathbf{H} + 6\mathbf{I} + 4\mathbf{J} + 3\mathbf{K} + \mathbf{L} + \mathbf{M}.$$

5. 如果向量 $\mathbf{K}_p$ 由 (4-21) 确定，使用问题 4 中的记号，证明

$$\mathbf{K}_2 = \mathbf{A} + hp_1\mathbf{B} + h^2\left[\frac{1}{2}p_1^2\mathbf{C} + p_1q_1\mathbf{D}\right]$$
$$+ h^3\left[\frac{1}{6}p_1^3\mathbf{E} + p_1^2q_1\mathbf{F}\right]$$
$$+ h^4\left[\frac{1}{24}p_1^4\mathbf{H} + \frac{1}{2}p_1^2q_1^2\mathbf{K} + \frac{1}{2}p_1^3q_1\mathbf{I}\right] + O(h^5).$$

验证由 (4-22) 确定的方法是形如 (4-21) 的唯一的方法，它对 $\kappa = 2$ 达到 $p = 3$。

6. 验证

$$\mathbf{K}_3 = \mathbf{A} + hp_2\mathbf{B} + h^2\left[\frac{1}{2}p_2^2\mathbf{C} + p_2q_2\mathbf{D}\right]$$
$$+ h^3\left[\frac{1}{6}p_2^3\mathbf{E} + p_2^2q_2\mathbf{F} + p_1p_2q_3\mathbf{G}\right]$$

$$+ h^4\left[\frac{1}{24}p_2^4\mathbf{H} + \frac{1}{2}p_2^3 q_2\mathbf{I} + \frac{1}{2}p_2^2 q_2^2\mathbf{K} + p_1 p_2^2 q_3\mathbf{J}\right.$$

$$\left.+ \frac{1}{2}p_1^2 p_2 q_3\mathbf{L} + p_1 p_2 q_1 q_3\mathbf{M}\right] + O(h^5).$$

7. 利用上面问题的结果，证明对于 $\kappa = 3$ 的方法 (4-21) 具有阶 $p = 4$，导出下面系数 a_i，b_i，p_i 及 q_i 值的单参数族：

$$p_1 = \frac{4t-1}{6t},\quad p_2 = \frac{2+t}{3},$$

$$q_1 = \frac{4t-1}{12t},\quad q_2 = \frac{2+t}{6},\quad q_3 = \frac{t(t+2t^2)}{2(4t-1)},$$

$$a_1 = \frac{2t-1+4t^3(1+t)}{(4t-1)(4+2t)(1+2t^2)},\ a_2 = \frac{t^2(1+2t)}{(4t-1)(1+2t^2)},$$

$$a_3 = \frac{1-t}{(4+2t)(1+2t^2)},$$

$$b_1 = \frac{2t-1+4t^3(1+t)}{(4t-1)(4+2t)(1+2t^2)},\ b_2 = \frac{6t^3}{(4t-1)(1+2t^2)},$$

$$b_3 = \frac{3}{(4+2t)(1+2t^2)}.$$

[对于 $t = 1$ 得到公式 (4-23)]

8*. 证明

$$\mathbf{K}_4 = \mathbf{A} + hp_3\mathbf{B} + h^2\left[\frac{1}{2}p_3^2\mathbf{C} + p_2 q_4\mathbf{D}\right]$$

$$+ h^3\left[\frac{1}{6}p_3^3\mathbf{E} + p_3^2 q_4\mathbf{F} + p_3(q_5 p_1 + q_6 p_2)\mathbf{G}\right]$$

$$+ h^4\left[\frac{1}{24}p_3^4\mathbf{H} + \frac{1}{2}p_3^3 q_4\mathbf{I} + p_3^2(q_5 p_1 + q_6 p_2)\mathbf{J}\right.$$

$$+ \frac{1}{2}p_3(q_5 p_1^2 + q_6 p_2^2)\mathbf{L} + p_3(p_1 q_1 q_5 + p_2 q_2 q_6)\mathbf{M}$$

$$\left.+ \frac{1}{2}p_3^2 q_4^2\mathbf{K}\right] + O(h^5).$$

并且验证 Nyström 公式 (4-24) 为 5 阶.

9. 对于形如

$$\boldsymbol{\Phi}(\mathbf{y},\mathbf{z};h)=\mathbf{z}+h(a_1\mathbf{K}_1+a_2\mathbf{K}_2),$$
$$\boldsymbol{\Psi}(\mathbf{y},\mathbf{z};h)=b_1\mathbf{K}_1+b_2\mathbf{K}_2$$

的特殊二阶方程确定所有三阶方法,其中

$$\mathbf{K}_1=\mathbf{f}(\mathbf{y}+hp_0\mathbf{z}),$$
$$\mathbf{K}_2=\mathbf{f}(\mathbf{y}+hp_1\mathbf{z}+h^2p_1q_1\mathbf{K}_1).$$

[公式 (4-22) 是 $p_0=0$ 的特殊情形.]

§4.3.

10. 对于三阶方法 (4-22),确定定理 4.2 中的常数 L 和 N 的界.

11*. 对于四阶方法 (4-23),重复问题 10。

12. 把方法 (4-22) 中的主误差函数的分量 $\boldsymbol{\varphi}$ 和 $\boldsymbol{\phi}$ 用问题 4 中所确定的向量 $\mathbf{A},\mathbf{B},\mathbf{C},\cdots$ 来表示,并验证关系式 (4-37).

13. 证明:对于方法 (4-23),主误差函数分量由

$$\boldsymbol{\varphi}(\mathbf{y},\mathbf{z})=-\frac{1}{720}\mathbf{E}-\frac{1}{240}\mathbf{F}-\frac{1}{120}\mathbf{G},$$

$$\boldsymbol{\phi}(\mathbf{y},\mathbf{z})=\frac{1}{2880}\mathbf{H}+\frac{11}{1920}\mathbf{I}+\frac{1}{120}\mathbf{J}+\frac{1}{720}\mathbf{K}+\frac{1}{480}\mathbf{L}+\frac{1}{480}\mathbf{M}$$

给出,并验证公式 (4-38).

14. 对问题 9 中构造的方法族,计算主误差函数 $\boldsymbol{\varphi}$,是否可用 $\boldsymbol{\phi}+\boldsymbol{\varphi}'=0$ 这种方式来选取参数? 这个结果的意义是什么?

15. 如果用三阶 Nyström 公式(4-23)解初值问题(3-99),试确定伸缩误差函数 $\mathbf{e}(x)$。 与三阶 Radau 公式比较其结果

怎样?

§4.4.

16. 研究初值问题 $y'' = y$, $y(0) = 1$, $y'(0) = -1$ 解的舍入误差传播，假设局部舍入误差为对称分布，并且 (a) 定点十进制运算; (b) 浮点十进制运算[在后一种情形，令 (4-62) 中 $c(x) = (y(x))^2$, $e(x) = (y'(x))^2$, $d(x) = 0$.]

17*. 以特殊的单步方法对微分方程 $y''' = f(x, y)$ 的数值积分以及对更一般的方程 $y^{(p)} = f(x, y)$ 阐述类似于 §§4.4-1 和 4.4-2 中所推导的那些结果.

注

§4.1. 化 n 阶方程为一阶方程组的优缺点由 Collatz 和 Zurmühl [1942] (也可见 Collatz [1960], 第11页), Rutishauser [1955], Ceschino 及 Kuntzmann [1960] 所讨论.

§4.2. 对于高阶方程的直接积分的其它方法，见 Nyström [1952], Babkin [1948], Rapoport [1952], de Vogelaere [1955], Varobev [1956] 以及在第 6 章中所讨论的方法.

第II部分　初值问题的多步方法

到现在为止，在我们所考虑过的方法中，在点 x_n 的近似解的增量 $y_{n+1}-y_n$ 是用一个仅仅依赖于（单个方程的情形）x_n，y_n 以及步长 h 的函数来计算的．这样得到的方法在概念上是简单的、通用的，并且容易编程序．另一方面，由于它们没有充分使用可以利用的信息，所以效率是低的．如果求 y_{n+1} 的值不仅依赖于 y_n 的值，譬如说还依赖于 y_{n-1} 和 y_{n-2}，似乎还有可能得到更加精确的值．根据这个想法所得到的一些方法已经变得很流行了． 在所要求的精度差不多的情形下，它们需要的工作一般比单步方法少．这个优点是以增加一些复杂性（需要一个特殊的初始处理）以及在某些情形下以离散误差和舍入误差（数值不稳定性）有增大的可能性为代价而获得的．

第五章　一阶方程的多步方法

在本章中我们将研究多步方法对初值问题

$$y' = f(x,y),\ \ y(a) = \eta \tag{5-1}$$

的应用，其中 $f(x,y)$ 满足第一章中存在定理的条件. 我们仅考虑含有单个微分方程的问题. 虽然所要讨论的方法本身就很容易地适用于求解微分方程组，但是在这种特殊情形中呈现了全部的基本性质.

5.1. 特殊的多步方法

5.1-1. *插值多项式*. 由于大多数常用的多步方法是以内插多项式为基础的，因此在这一节中我们将回顾 Lagrange 插值多项式的一些性质. 设函数 $z(x)$ 在含有 $q+1$ 个不同点 $x_0, x_1, \cdots, x_q$ 的区间 J 上是确定的. 众所周知，在所有次数不超过 q 的 x 的多项式中，恰好存在一个多项式 $P(x)$，它满足关系式

$$P(x_\nu) = z(x_\nu),\ \ \nu = 0,1,\cdots,q;$$

或者换句话说，在点 x_ν 处插值函数 $z(x)$. 这个称为内插多项式的唯一性是根据任何两个这种多项式的差是一个次数 $\leqslant q$ 且具有 $q+1$ 个零点(即在点 x_ν 处)的多项式，从而恒为零的事实而得到的. 通过多项式的显式表示，可以证明它的存在性. 例如其形式为

$$P(x) = L(x)\sum_{\nu=0}^{q}\frac{z(x_\nu)}{L'(x_\nu)(x-x_\nu)}, \tag{5-2}$$

其中

$$L(x)=(x-x_0)(x-x_1)\cdots(x-x_q). \qquad (5-3)$$

自然希望，$P(x)$不仅在点 $x=x_\nu$ 处可精确地表达出 $z(x)$，而且对 x 的其它值也可近似地表达出 $z(x)$. 下面的引理可估计出由这种近似所引起的误差.

引理 5.1. 令 $z(x)$ 在 J 内有 $q+1$ 阶连续导数，则对 J 内的每一点 x，在包含 x 和点 $x_\nu(\nu=0,\cdots,q)$ 的最小区间内存在一点 ξ，使得

$$z(x)-P(x)=\frac{1}{(q+1)!}L(x)z^{(q+1)}(\xi). \qquad (5-4)$$

证. 若 x 是点 x_ν 中的一个点便无需证明，因为(5-4)的两端对任意 ξ 全为零. 如果 $x\neq x_\nu$, $\nu=0,\cdots,q$，考虑变量 t 的辅助函数

$$Z(t)=z(t)-P(t)-\lambda L(t),$$

其中

$$\lambda=\frac{z(x)-P(x)}{L(x)}. \qquad (5-5)$$

因此 $Z(x)=0$. 显然函数 $Z(t)$ 在所有点 x_ν 处也为零，因而它在区间 I 的 $q+2$ 个不同点上等于零. 重复应用 Rolle 定理，于是它的 μ 阶导数至少在 I 的 $q+2-\mu$ 个不同点上为零 ($\mu=1,2,\cdots,q+1$). 现在令 ξ 是使其 $q+1$ 阶导数为零的点. 因为 $P^{(q+1)}(\xi)\equiv 0$ 且 $L^{(q+1)}(\xi)=(q+1)!$，所以在这一点上就有

$$Z^{(q+1)}(\xi)=z^{(q+1)}(\xi)-\lambda(q+1)!=0.$$

利用(5-5)并移项后，便得我们所需要的结果.

注 虽然量 ξ 作为 x 的函数一般来说是未知的，但是我们可以断定 $z^{(q+1)}(\xi)$ 为 x 的连续函数，$x\in J$. 这由(5-4)解出 $z^{(q+1)}(\xi)$ 便容易看出. 于是，对于 $x\neq x_\nu$，$z^{(q+1)}(\xi)$ 显然是

连续的;而在点 $x=x_\nu$ 处其连续性可以根据 L'Hospital 法则得到,这是因为 $L'(x_\nu)=\prod\limits_{\mu\neq\nu}(x_\mu-x_\nu)\neq 0$.

我们还将用 $P(x)$ 的导数来近似 $z'(x)$. 在点 x_ν 处这个近似的量由下述结果[1]以一个简单方法估计出来的.

引理 5.2. 令 $z(x)$ 满足引理 5.1 中相同的假设，则对每一个 $x_\mu(\mu=0,1,\cdots,q)$ 存在一个数 $\xi\in J$，使得

$$z'(x_\mu)-P'(x_\mu)=\frac{1}{(q+1)!}z^{(q+1)}(\xi)L'(x_\mu). \qquad (5-6)$$

证. 不失一般性，我们可以假设 $x_0<x_1<\cdots<x_q$. 对任选的常数 λ，函数

$$Z(x)=z(x)-P(x)-\lambda L(x)$$

有 $q+1$ 个零点 $x=x_\nu$, $\nu=0,\cdots,q$. 利用 Rolle 定理，$Z'(x)$ 有 q 个零点，即在每个区间 $(x_\nu,x_{\nu+1})$ $(\nu=0,\cdots q-1)$ 内有一个零点. 现选取 λ 使得 $Z'(x)$ 在 $x=x_\mu$ 处有另一个零点,这就导出条件

$$\lambda=\frac{z'(x_\mu)-P'(x_\mu)}{L'(x_\mu)}. \qquad (5-7)$$

由于 $L'(x_\mu)\neq 0$，它是成立的. 现在函数 $Z'(x)$ 在 J 内有 $q+1$ 个零点,重复应用 Rolle 定理就得到 $Z^{(q+1)}(x)$ 在 J 内至少有一个零点,譬如说 $x=\xi$. 于是,由 (5-7) 便得到结论

$$Z^{(q+1)}(\xi)=z^{(q+1)}(\xi)-\lambda(q+1)!.$$

在应用于微分方程时,点 x_ν 通常是等距的. 因此用有限差分来表示内插多项式是非常方便的. 但是还存在一件使人为难的事情，即有限差分符号有各种定义以及由它们可产生出各种公式[2]. 而为了形式上简便，我们只采用一种差分符

---

1) 在任意点 x 处的估计是更为复杂的(见 Hildebran [1956] p.66).

2) 例如,见 Hildebrand [1956]，第四章和第五章.

号．这是因为内插多项式本身是唯一的，只不过所能表达的各种形式是不同的，所以，这样做除了外形上可能欠工整外并不受影响．

我们假定 $x_p = x_0 + ph$，这里 h 是一个常数，p 是一个整数，并记 $z_p = z(x_p)$．在点 $x = x_p$ 处函数 $z(x)$ 的一阶向后差分规定为

$$\nabla z_p = z_p - z_{p-1}, \tag{5-8}$$

更高阶的向后差分规定为 $\nabla^q z_p = \nabla(\nabla^{q-1} z_p)$ 例如，

$$\begin{aligned}\nabla^2 z_p &= (z_p - z_{p-1}) - (z_{p-1} - z_{p-2}) \\ &= z_p - 2z_{p-1} + z_{p-2}.\end{aligned}$$

为了对称起见，令 $\nabla^0 z_p = z_p$．利用归纳法容易证明

$$\nabla^q z_p = \sum_{m=0}^{q} (-1)^m \binom{q}{m} z_{p-m}, \quad q = 0, 1, \cdots, \tag{5-9}$$

其中 $\binom{q}{m}$ 表示二项式系数

$$\begin{aligned}&\binom{q}{0} = 1, \\ &\binom{q}{m} = \frac{q(q-1)\cdots(q-m+1)}{1 \cdot 2 \cdots m}, \quad m = 1, 2, \cdots.\end{aligned} \tag{5-10}$$

当 q 不是自然数时它们也有定义．

公式 (5-9) 用函数值或纵坐标表示差分，有时也需要用差分来表示纵坐标．我们有

$$\begin{aligned}z_p &= \nabla^0 z_p, \\ z_{p-1} &= z_p - (z_p - z_{p-1}) = \nabla^0 z_p - \nabla^1 z_p, \\ z_{p-2} &= z_p - 2(z_p - z_{p-1}) + (z_p - 2z_{p-1} + z_{p-2}) \\ &= \nabla^0 z_p - 2\nabla^1 z_p + \nabla^2 z_p,\end{aligned}$$

用数学归纳法容易证明下述公式成立：

$$z_{p-q}=\sum_{m=0}^{q}(-1)^m\binom{q}{m}\nabla^m z_p,\quad q=0,1,\cdots.\qquad(5\text{-}11)$$

为了用差分来表示插值多项式，宜于引进变量

$$s=\frac{x-x_p}{h},$$

它表示从 $x=x_p$ 出发以 h 为单位来度量 x. 容易证明在点 $x_p,x_{p-1},\cdots,x_{p-q}$ 插值函数 $z(x)$ 的次数 $\leqslant q$ 的多项式 $P(x)$ 可表示成

$$P(x)=\sum_{m=0}^{q}(-1)^m\binom{-s}{m}\nabla^m z_p.\qquad(5\text{-}12)$$

因为 $\binom{-s}{m}$ 是 s 的 m 次多项式，又因为 s 是 x 的线性函数，因此 (5-12) 确实是次数 $\leqslant q$ 的多项式. 而且，由于当整数 $r\leqslant m$ 时，$\binom{r}{m}=0$，所以由 (5-11)，对于 $r=0,1,\cdots,q$，有

$$P(x_{p-r})=\sum_{m=0}^{q}(-1)^m\binom{r}{m}\nabla^m z_p$$

$$=\sum_{m=0}^{r}(-1)^m\binom{r}{m}\nabla^m z_p=z_{p-r}.$$

内插多项式的表达式 (5-12) 称为 Newton 向后差分公式，并在以下的章节中常常采用它.

5.1-2. **建立在数值积分基础上的方法.** 按定义，微分方程 (5-1) 的精确解对于区间 $[a,b]$ 内的任意二点 x 和 $x+k$ 都满足恒等式

$$y(x+k)-y(x)=\int_x^{x+k}f(t,y(t))dt.\qquad(5\text{-}13)$$

现在所要讨论的方法是以内插多项式来代替未知函数 $f(x,$

$y(x)$）为基础的，这个内插多项式在一系列点 x_n 上取值

$$f_n = f(x_n, y_n),$$

其中 y_n 是已经计算出或正要计算出的量，计算出这个积分并采用它的值作为近似值 y_n 在 x 和 $x+k$ 之间的增量。我们假设内插点为 $x_p, x_{p-1}, \cdots, x_{p-q}$，那么代替 $f(x, y(x))$ 的多项式由

$$P(x) = \sum_{m=0}^{q} (-1)^m \binom{-s}{m} \nabla^m f_p, \quad s = \frac{x - x_p}{h}$$

给出。

在原则上正整数 q 是任意的，而实际上却很少超过 6，按照 x 和 $x+k$ 相对于插值点的位置，可分成几类方法，见表 5.1。我们将比较详细地考察其中的每一个方法。

表 5.1

方　　法	x	$x+k$
Adams-Bashforth	x_p	x_{p+1}
Adams-Moulton	x_{p-1}	x_p
Nyström	x_{p-1}	x_{p+1}
Milne-Simpson	x_{p-2}	x_p

(i) Adams-Bashforth 方法（Bashforth and Adams[1883]）。此时我们有

$$y_{p+1} - y_p = \int_{x_p}^{x_{p+1}} P(x)dx = h \sum_{m=0}^{q} \gamma_m \nabla^m f_p, \quad (5\text{-}14)$$

其中常数

$$\gamma_m = (-1)^m \frac{1}{h} \int_{x_p}^{x_{p+1}} \binom{-s}{m} dx = (-1)^m \int_0^1 \binom{-s}{m} ds \quad (5\text{-}15)$$

是与 f 无关的，并且下面将进行数值计算。若值 $y_p, y_{p-1}, \cdots,$

y_{p-q} 为已知，则可以求得对应于 $f_n = f(x_n, y_n)$ 的值，并且容易形成差分 $\nabla^m f_p$. 从而(5-14)式的右端表达式是已知的，于是 $y_{p+1} - y_p$ 可以计算出来，因此可以求得 y_{p+1} 的值. 然后将下标 p 增加 1，再用同样的公式计算出 y_{p+2}，等等. 如果 $y_p, y_{p-1}, \cdots, y_{p-q}$ 中有些是未知的，则该方法就失效. 通常开始计算时初始条件只提供了所要的 $q+1$ 个值中的一个值或在改变步长时便是这种情形. 在这种情形下，所缺少的值必须用一个单独的方法来求得. 例如象 Runge-Kutta 那样的单步方法或 Taylor 展式都是常用的方法. 属于差分法范围内的其它一些起步方法也是熟知的 (Collatz [1960], p.81). 缺少起步值的困难对所有多步法都是有代表性的.

为了求得系数 γ_m 的递推关系，数值和其它有用的性质，我们采用生成函数的方法. 假设这个函数为 $G(t)$，按其 Maclaurin 展式以 γ_m 为系数. 利用二项式定理的一般形式，我们得到

$$G(t) = \sum_{m=0}^{\infty} \gamma_m t^m = \sum_{m=0}^{\infty} (-t)^m \int_0^1 \binom{-s}{m} ds$$

$$= \int_0^1 \sum_{m=0}^{\infty} (-t)^m \binom{-s}{m} ds = \int_0^1 (1-t)^{-s} ds.$$

令 $(1-t)^{-s} = e^{-s\log(1-t)}$，容易求得该积分. 由此我们得到[1]

$$G(t) = -\frac{t}{(1-t)\log(1-t)}, \qquad (5-16)$$

它又可写成

$$-\frac{\log(1-t)}{t} G(t) = \frac{1}{1-t}.$$

利用熟知的展式

1) 这种形式上的运算是根据$G(t)$是复变量 t 的解析函数来保证的，$|t| < 1$.

$$\frac{1}{1-t} = 1 + t + t^2 + \cdots,$$
$$-\frac{\log(1-t)}{t} = 1 + \frac{1}{2}t + \frac{1}{3}t^2 + \cdots, \tag{5-17}$$

我们得到幂级数之间的如下恒等式:

$$\left(1 + \frac{1}{2}t + \frac{1}{3}t^2 + \cdots\right)(\gamma_0 + \gamma_1 t + \gamma_2 t^2 + \cdots)$$
$$= 1 + t + t^2 + \cdots.$$

比较对应于 t 的各次幂的系数,我们得到关系式

$$\gamma_m + \frac{1}{2}\gamma_{m-1} + \frac{1}{3}\gamma_{m-2} + \cdots + \frac{1}{m+1}\gamma_0 = 1,$$
$$m = 0, 1, 2, \cdots.$$

由这个关系式可以递推计算出 γ_m. 表 5.2 给出的值就是用这种方法求得的.

表 5.2

m	0	1	2	3	4	5	6
γ_m	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19087}{60480}$

明显采用差分 $\nabla^m f_p$ 表示的公式(5-14),特别适于手算.因为在计算中随时可以发挥人的主动性和判断力,从而改变计算过程. 例如,如果最高阶差分 $\nabla^q f_p$ 值较大,则可看成是精确度不够并要增加(5-14)的项数的一个标志. 但是,如果一旦固定了数 q,通常为计算机编制的 Adams 方法便是这样,就没有特别的理由一定要采用差分表示. 由(5-9)用纵坐标来表示差分并且合并相同纵坐标的系数,Adams-Bashforth 公式以形式

$$y_{p+1} - y_p = h\sum_{\rho=0}^{q} \beta_{q\rho} f_{p-\rho} \tag{5-18}$$

出现，其中系数 β_{qp} 由下式给出：

$$\beta_{qp}=(-1)^{p}\left\{\binom{p}{p}\gamma_{p}+\binom{p+1}{p}\gamma_{p+1}+\cdots+\binom{q}{p}\gamma_{q}\right\}$$

$$p=0,1,\cdots,q;\quad q=0,1,\cdots.$$

应当注意到系数 β_{qp} 依赖于 q 和 p，这就使得改变所用差分的阶数更加困难．表 5.3 给出了一些数值．这些系数的数值较大而且符号交替出现是该方法的一个不利条件，见 §5.3-4．

表 5.3　系数 β_{qp}

p	0	1	2	3	4	5
β_{0p}	1					
$2\beta_{1p}$	3	−1				
$12\beta_{2p}$	23	−16	5			
$24\beta_{3p}$	55	−59	37	−9		
$720\beta_{4p}$	1901	−2774	2616	−1274	251	
$1440\beta_{5p}$	4277	−7923	9982	−7298	2877	−475

ii) Adams-Moulton 方法 (Moulton [1926])．现在对于方程 (5-13) 采用如下形式：

$$y_{p}-y_{p-1}=\int_{x_{p-1}}^{x_{p}}P(x)dx=h\sum_{m=0}^{q}\gamma_{m}^{*}\nabla^{m}f_{p},\quad(5\text{-}19)$$

其中

$$\gamma_{m}^{*}=(-1)^{m}h^{-1}\int_{x_{p-1}}^{x_{p}}\binom{-s}{m}dx=(-1)^{m}\int_{-1}^{0}\binom{-s}{m}ds.\quad(5\text{-}20)$$

通过类似于导出 (5-16) 的那些运算，系数 γ_{m}^{*} 的生成函数确定如下：

$$G^{*}(t)=\sum_{m=0}^{\infty}\gamma_{m}^{*}t^{m}=-\frac{t}{\log(1-t)},\quad(5\text{-}21)$$

故有

$$-\frac{\log(1-t)}{t}G^*(t)=1.$$

或利用展式(5-17),

$$\left(1+\frac{1}{2}t+\frac{1}{3}t^2+\cdots\right)(\gamma_0^*+\gamma_1^*t+\gamma_m^*t^2+\cdots)=1.$$

从而得到

$$\gamma_m^*+\frac{1}{2}\gamma_{m-1}^*+\frac{1}{3}\gamma_{m-2}^*+\cdots+\frac{1}{m+1}\gamma_0^*$$

$$=\begin{cases}1, & m=0,\\ 0, & m=1,2,3,\cdots.\end{cases}$$

由这些递推关系容易得到表(5.4)中给出的数值.

表 5.4

m	0	1	2	3	4	5	6
γ_m^*	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$	$-\frac{863}{60480}$

我们还注意到关系式

$$\frac{1}{1-t}G^*(t)=G(t)$$

或

$$(1+t+t^2+\cdots)(\gamma_0^*+\gamma_1^*t+\cdots)$$
$$=\gamma_0+\gamma_1t+\gamma_2t^2+\cdots,$$

比较 t^m 的系数,我们得到

$$\gamma_0^*+\gamma_1^*+\cdots+\gamma_m^*=\gamma_m,\quad m=0,1,2,\cdots. \tag{5-22}$$

公式(5-19)和 Adams-Bashforth 公式一样使用,只是现在仅仅知道值 $y_{p-1}, y_{p-2}, \cdots, y_{p-q}$ 而由(5-19)确定y_p. 由于 y_p 作为 $f_p=f(x_p, y_p)$ 的自变量出现在(5-19)的右端项中,因此现在这个方程是 y_p 的一个非显式方程. 一般地说它不可能显式地解出. 幸而是该方程的特殊形式提供了一个迭

代过程，当 h 充分小时，它很快地给出解。

假定由某种来源已经得到 (5-19) 解的一个近似值 $y_p^{(0)}$，我们就可求出 $f_p^{(0)}=f(x_p, y_p^{(0)})$ 并且建立差分 $\nabla f_p^{(0)}=f_p^{(0)}-f_{p-1}$，$\nabla^2 f_p^{(0)}=\nabla f_p^{(0)}-\nabla f_{p-1}\cdots$。然后由下式得到一个较好的近似值 $y_p^{(1)}$

$$y_p^{(1)}=y_{p-1}+h\sum_{m=0}^{q}\gamma_m^*\nabla^m f_p^{(0)}. \qquad (5\text{-}23a)$$

求出 $f_p^{(1)}=f(x_p, y_p^{(1)})$ 后，重新计算差分，则可以得到更好些的近似值 $y_p^{(2)}$

$$y_p^{(2)}=y_{p-1}+h\sum_{m=0}^{q}\gamma_m^*\nabla^m f_p^{(1)}. \qquad (5\text{-}23b)$$

一般地说，近似值序列 $y_p^{(\nu)}(\nu=0,1,2,\cdots)$ 由关系式

$$y_p^{(\nu+1)}=y_{p-1}+h\sum_{m=0}^{q}\gamma_m^*\nabla^m f_p^{(\nu)} \qquad (5\text{-}23c)$$

递推地得到，其中 $f_p^{(\nu)}=f(x_p, y_p^{(\nu)})$。根据 §5.2-2 中将要证明的定理 5.4 推得，这样确定的数列 $y_p^{(0)}, y_p^{(1)}, y_p^{(2)}, \cdots$ 对充分小的 h 值收敛于 (5-19) 的一个解，并且这个解是唯一的。

实际数值计算时并不需要[1]在每个迭代步计算整个公式 (5-23c)。由关系式 (5-23c) 减去用 $\nu-1$ 代替 ν 后得到的相应的关系式，我们得到

$$y_p^{(\nu+1)}-y_p^{(\nu)}=h\sum_{m=0}^{q}\gamma_m^*(\nabla^m f_p^{(\nu)}-\nabla^m f_p^{(\nu-1)}).$$

差分 $\nabla^m f_p^{(\nu)}$ 和 $\nabla^m f_p^{(\nu-1)}$ 只在最前面的纵坐标 $f_p^{(\nu)}$ 与 $f_p^{(\nu-1)}$ 上不同，从而推得

$$\nabla^m f_p^{(\nu)}-\nabla^m f_p^{(\nu-1)}=f_p^{(\nu)}-f_p^{(\nu-1)},$$

所以

1) 下面的简化是由 Stohler [1943] 得到的。

$$y_p^{(\nu+1)} - y_p^{(\nu)} = h\sum_{m=0}^{q}\gamma_m^*(f_p^{(\nu)} - f_p^{(\nu-1)}).$$

利用(5-22)，有

$$y_p^{(\nu+1)} - y^{(\nu)} = h\gamma_q(f_p^{(\nu)} - f_p^{(\nu-1)}). \qquad (5\text{-}23d)$$

现在我们通过对级数

$$y_p^* = y_p^{(1)} + (y_p^{(2)} - y_p^{(1)}) + (y_p^{(3)} - y_p^{(2)}) + \cdots$$

的项求和来获得(5-19)的解 y_p，该级数由重复使用(5-23d)而得到. 因此在实际计算时，只要当 $f_p^{(\nu)} - f_p^{(\nu-1)}$ 可以忽略不计时，便可停止计算，而把最后一次得到的 $f_p^{(\nu)}$ 的数值就作为 f_p 的终值. 由它再形成差分 $\nabla^m f_p$ 的终值. 因为通过求级数和得到的 y_p^* 的数值可能会受到舍入误差的影响，所以最好重新用公式(5-19)计算出 y_p 的终值.

上面讲到的为迭代过程提供首次近似值 $y_p^{(0)}$ 的公式就叫做预估公式，而公式(5-23)就称为校正公式. 于是上述迭代过程就是先预估一个 y_p 的试探性值 $y_p^{(0)}$，然后通过校正公式来校正它(可能需要若干次). 为了使校正次数减到最少，显然需要预估值 $y_p^{(0)}$ 尽可能精确. 如果预估公式相当精确，那么就有可能从预估值 $y_p^{(0)}$ 与终值 y_p 之差得出局部离散误差的结论(见§5.3-6). 而且，一个充分精确的预估公式无需进行多次校正(见§5.3-7). 由于这个原因，Adams-Bashforth 公式[公式(5-14)中 p 减去 1]被推荐为 Adams-Moulton 方法的一个精确的预估公式.

作为一个数值例子，我们在(5-19)中用 $q = 2$, $h = 0.1$ 的 Adams-Moulton 方法完成初值问题:

$$y' = x - y^2, \quad y(0) = 0$$

解的开头几步积分，此时公式写成

$$y_p - y_{p-1} = h\left\{f_p - \frac{1}{2}\nabla f_p - \frac{1}{12}\nabla^2 f_p\right\},$$

而用 Adams-Bashforth 公式

$$y_p - y_{p-1} = h\left\{f_{p-1} + \frac{1}{2}\nabla f_{p-1} + \frac{5}{12}\nabla^2 f_{p-1}\right\}$$

作为预估公式。所需要的三个开始值可以从精确解的级数展开式

$$y(x) = \frac{1}{2}x^2 - \frac{1}{20}x^5 + \frac{1}{160}x^8 - \cdots$$

得到。开始值的横坐标取成关于"准确"初值 $x_0 = 0$ 的对称点 $x = -h$ 和 $x = h$，在 $x = -h$ 及 $x = h$ 处求出这个级数的值。

现在我们把对应于 x_{-1}，x_0 和 x_1 的数值填入表 5.5 中，剩下的工作是按上述过程进行。表中箭头表示计算这些数值的顺序；箭头 ↙——表示用该箭头上面一行及其右边的全部数值计算左边的结果。对 $x = x_3$ 仅仅以简略的形式填在表中。

用纵坐标表示差分，Adams-Moulton 公式 (5-19) 可以写成形式

$$y_p - y_{p-1} = h\sum_{\rho=0}^{q}\beta_{q\rho}^{*} f_{p-\rho}, \tag{5-24}$$

其中

$$\beta_{q\rho}^{*} = (-1)^{\rho}\left\{\binom{\rho}{\rho}\gamma_{\rho}^{*} + \binom{\rho+1}{\rho}\gamma_{\rho+1}^{*} + \cdots\binom{q}{\rho}\gamma_{q}^{*}\right\},$$

$$\rho = 0, 1, \cdots, q;\quad q = 0, 1, \cdots.$$

表 5.6 中给出 $\beta_{q\rho}^{*}$ 的某些数值。

图 5.1 是 Adams-Moulton 方法对初值问题求解的框图，它是用 Adams-Bashforth 公式作预估公式并用纵坐标进行计算的。假定开始值 $y_0, y_1, \cdots, y_{q-1}$ 是由增量函数 Φ 所确定的单步法产生的。

由于确定新值 y_p 需要解一个方程，因此 Adams-Moulton

表 5.5 Adams-Moulton 方法的例子

x_n	y_n	f_n	∇f_n	$\nabla^2 f_n$
−0.1	0.0050005000	−0.1000250050		
0	0	0	0.1000250050	
0.1	0.0049995000	0.0999750050	0.0999750050	−0.000050000
0.2	$y_2^{(0)}=0.0199936674$	$f_2^{(0)}=0.1996002533$	0.0996252483	−0.0003497567
	$y_2^{(1)}=0.0199811776$	$f_2^{(1)}=0.1996007525$	用 $f(x_2,y_2^{(0)})$ 形成的试探性差分	
	$y_2^{(2)}-y_2^{(1)}=$ 208	$\leftarrow f_2^{(1)}-f_2^{(0)}=$ 4992		
	$y_2^{(2)}=0.0199811984$	$f_2^{(2)}=0.1996007517$		
	−0	$\leftarrow f_2^{(2)}-f_2^{(1)}=$ −8		
	$y_2^*=y_2^{(3)}=0.0199811984$	$f_2^{(3)}=0.1996007517$	0.0996257467	−0.0003492583
	$y_2=0.0199811983$			
0.3	$y_3^{(0)}=0.0449080084$	$f_3^{(0)}=0.2979832708$	0.0983825919	−0.0012432276
			试探性差分	
	$y_3^{(1)}=0.0448707597$	$f_3^{(1)}=0.2979866149$		
	$y_3^{(1)}=0.0448708990$	$f_3^{(2)}=0.2979866024$		
	$y_3^*=y_3^{(3)}=0.0448708985$	$f_3^{(3)}=0.2979866024$	0.0983858507	−0.0012398960
	$y_3=0.0448708985$			

方法称为隐式方法，与此对比，Adams-Bashforth 方法称为显式方法。 Adams-Moulton 公式计算上的复杂性还是可以接受的，因为它可以得到更为精确的结果。显然，这是因为它使用了更精确的插值多项式；见 §5.1-3 的正式讨论。 Adams-Moulton 方法的特殊情形 $q=1$ 就是我们熟悉的梯形法。

表 5.6 系数 β^*_{qp}

p	0	1	2	3	4	5
β^*_{0p}	1					
$2\beta^*_{1p}$	1	1				
$12\beta^*_{2p}$	5	8	−1			
$24\beta^*_{3p}$	9	19	−5	1		
$720\beta^*_{4p}$	251	646	−264	106	−19	
$1440\beta^*_{5p}$	475	1427	−798	482	−173	27

iii) Nyström 方法 (Nyström [1925])。在这里我们令

$$y_{p+1}-y_{p-1}=\int_{x_{p-1}}^{x_{p+1}} P(x)\,dx=h\sum_{m=0}^{q}\kappa_m\nabla^m f_p, \quad (5\text{-}25)$$

其中

$$\kappa_m=(-1)^m\frac{1}{h}\int_{x_{p-1}}^{x_{p+1}}\binom{-s}{m}dx=(-1)^m\int_{-1}^{1}\binom{-s}{m}ds. \quad (5\text{-}26)$$

这是一个显式方法，具体算法与 Adams-Bashforth 方法非常相似，只是现在计算 y_n 的增量是用二步代替了一步。这个方法仅是弱稳定的，所有计算 y_n-y_m 的方法都具有这一共同性质[1]，其中 $n-m>1$。

系数 κ_m 的生成函数是

$$K(t)=\sum_{m=0}^{\infty}\kappa_m t^m=-\frac{t}{\log(1-t)}\cdot\frac{2-t}{1-t}. \quad (5\text{-}27)$$

1) 弱稳定现象将在 §5.3 和 §5.4 中讨论。

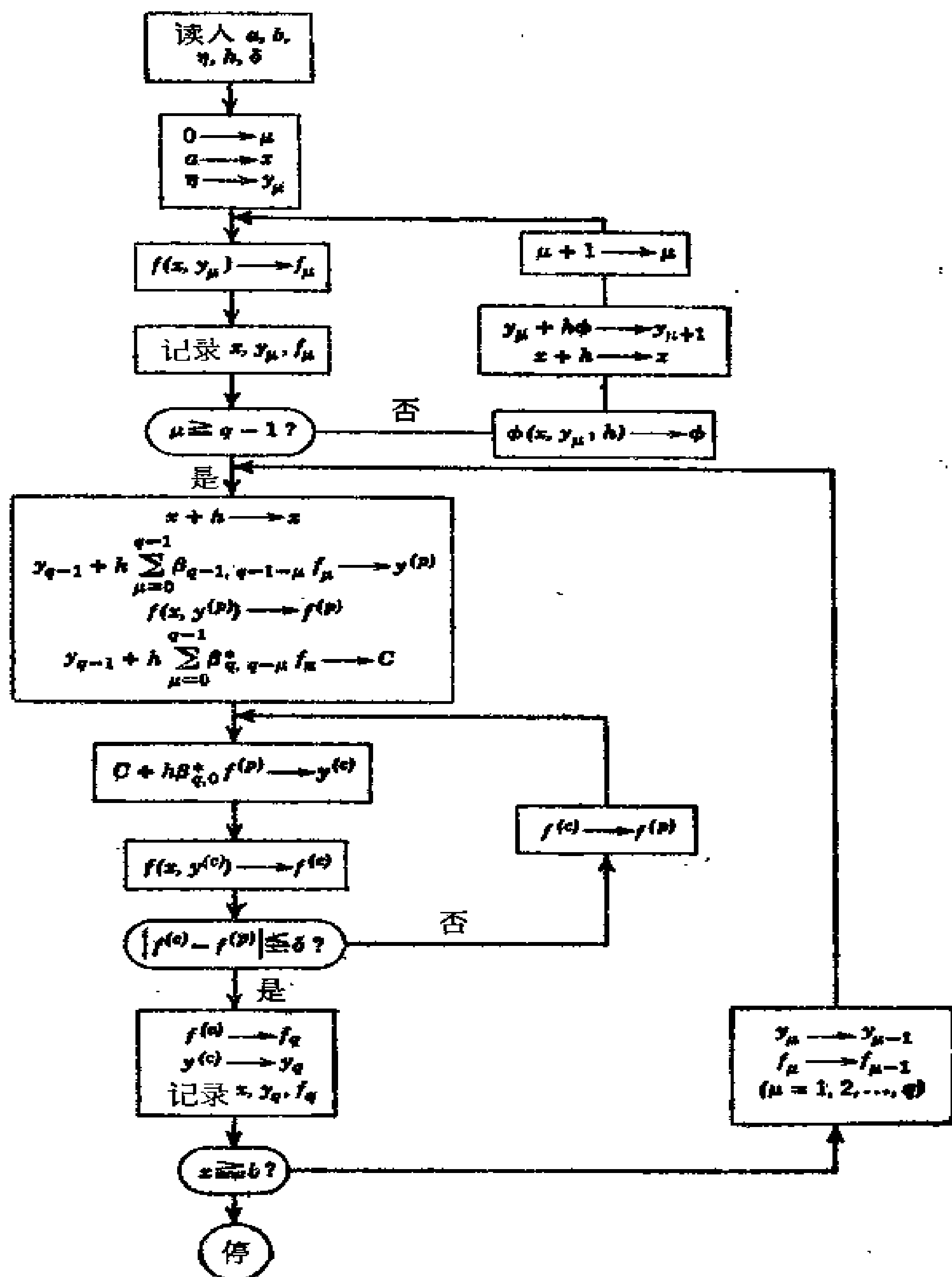


图 5.1 Adams-Moulton 方法的框图

利用展式(5-17)及

$$\frac{2-t}{1-t}=1+\frac{1}{1-t}=2+t+t^2+\cdots,$$

我们就得到恒等式

$$\left(1+\frac{1}{2}t+\frac{1}{3}t^2+\cdots\right)(\kappa_0+\kappa_1 t+\kappa_2 t^2+\cdots)$$
$$=2+t+t^2+\cdots$$

由此我们可以导出递推关系

$$\kappa_m+\frac{1}{2}\kappa_{m-1}+\frac{1}{3}\kappa_{m-2}+\cdots+\frac{1}{m+1}\kappa_0$$
$$=\begin{cases}2, & m=0,\\ 1, & m=1,2,\cdots.\end{cases}$$

表 5.7 中给出得到的一些数值.

表 5.7

m	0	1	2	3	4	5	6
κ_m	2	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{29}{90}$	$\frac{14}{45}$	$\frac{1139}{3780}$

Nyström 方法当 $q=0$ 时的特殊情形就称为中点法则. 因为 $\kappa_1=0$, 所以它给出通常情况下取 $q=1$ 时可期望得到的精确度,这是该方法常用的一个原因.中点法则也曾用来作为梯形法 (Adams-Moulton 取 $q=1$) 的预估公式. 对于一般的 q 而言, Nyström 方法可以作为下面即将讨论的推广的 Milne-Simpson 方法的预估公式.

iv) 推广的 Milne-Simpson 方法. 在公式(5-13)中取 $x=x_{p-2}$, $\kappa=2h$ 便得这个公式. 于是我们有公式

$$y_p-y_{p-2}=h\int_{x_{p-2}}^{x_p}P(x)dx=h\sum_{m=0}^{q}\kappa_m^*\nabla^m f_p, \quad (5\text{-}28)$$

其中

$$\kappa_m^* = (-1)^m \frac{1}{h}\int_{x_{p-1}}^{x_p}\binom{-s}{m}dx = (-1)^m\int_{-2}^{0}\binom{-s}{m}ds. \tag{5-29}$$

公式(5-28)类似 Adams-Moulton 方法,它是隐式方法.这个积分越过了二个步长,这种情形可能导致弱稳定.另一方面,对于所要求的精确度差不多的情形 $q \geqslant 2$ 的方法要比以前所讨论过的任何方法更为精确.特别是对于 $q = 2$ 的情形,这就是大家知道的 Milne 方法 (Milne [1926], [1953], p.66).确切地说,该结论也是成立的. 其原因可以通过研究系数 κ_m^* 来理解.它们的生成函数为

$$K^*(t) = \sum_{m=0}^{\infty}\kappa_m^* t^m = -\frac{t}{\log(1-t)}(2-t). \tag{5-30}$$

从而

$$\left(1+\frac{1}{2}t+\frac{1}{3}t^2+\cdots\right)(\kappa_0^*+\kappa_1^*t+\kappa_2^*t^2+\cdots)$$
$$=2-t.$$

比较系数,我们得到

$$\kappa_m^*+\frac{1}{2}\kappa_{m-1}^*+\frac{1}{3}\kappa_{m-2}^*+\cdots+\frac{1}{m+1}\kappa_0^*$$
$$=\begin{cases}2, & m=0,\\ -1, & m=1,\\ 0, & m=2,3,\cdots.\end{cases}$$

在表(5.8)中给出 κ_m^* 的一些数值. 我们注意到 $\kappa_3^* = 0$, 这

表 5.8

m	0	1	2	3	4	5	6
κ_m^*	2	-2	$\frac{1}{3}$	0	$-\frac{1}{90}$	$-\frac{1}{90}$	$-\frac{37}{3780}$

表明 $q=2$ 时 Milne-Simpson 公式给出通常在 $q=3$ 时才能期望的结果．我们还注意到

$$\frac{1}{1-t}K^*(t)=K(t),$$

由此可得关系式

$$\kappa_0^*+\kappa_1^*+\kappa_2^*+\cdots+\kappa_m^*=\kappa_m,\quad m=0,1,\cdots. \tag{5-31}$$

表 5.8 中给出的数值提示我们把公式(5-28)可写成便于使用的形式

$$y_p-y_{p-2}=h\left[2f_{p-1}+\frac{1}{3}\nabla^2 f_p-\frac{1}{90}(\nabla^4 f_p+\nabla^5 f_p)+\cdots\right], \tag{5-32}$$

从 $q=2$ 或 3 所得到的 Milne 公式可用纵坐标的形式写成

$$y_p-y_{p-2}=\frac{1}{3}h(f_p+4f_{p-1}+f_{p-2}). \tag{5-33}$$

如果 f 仅仅与 x 有关，这便化成数值积分理论中所熟悉的 Simpson 公式．

方程(5-28)，(5-32)和(5-33)都是关于 y_p 的隐式方程，并且通常是从一个预估的首次近似值 $y_p^{(0)}$ 起步，使用迭代法求解．原则上任何一个显式公式都可以用来计算 $y_p^{(0)}$．但是为了减少校正次数，如果有可能的话，总是选取一个其精确度与校正公式相当的预估公式．Milne 建议采用公式

$$y_p-y_{p-4}=\frac{1}{3}h[8f_{p-1}-4f_{p-2}+8f_{p-3}] \tag{5-34}$$

作为预估公式[从 x_{p-4} 到 x_p 之间积分通过点 x_{p-1}，x_{p-2}，x_{p-3} 处插值 $f(x,y(x))$ 的二次多项式而得到]．由于公式(5-31)，利用

$$y_p-y_{p-2}=h\kappa_m f(x_p,y_p)+f_{p-1},\cdots,f_{p-q}$$

的事实，对任意选取的预估公式，关于 Adams-Moulton 公式，按照本节所阐明的方法可以进行这个迭代过程。

5.1-3. 建立在积分基础上的公式的局部误差。 前一节我们遇到了许多差分表达式，它们近似一个依赖于给定函数 $f(x)$ 的量（即定积分的值）。这个量就称为泛函。我们用符号 Lf 来表示它。这种以离散算子 $L_h f$ 来近似 Lf 的泛函问题在数值计算中经常出现。度量这种有限差分近似值的精确度可以通过用 L_h 来代替 L 对函数的作用，并根据 h 和函数 f 的性质来度量偏差

$$R = Lf - L_h f.$$

量 R 可以看成相当于单步法中所讨论过的局部离散误差。在 §5.3-4 中我们将看到 R 的大小是怎样影响累积舍入误差的。

对于 Adams-Bashforth 和 Adams-Moulton 方法，R 的表达式可通过 $y'(x)=f(x,y(x))$ 的内插多项式加上余项来表示可以容易得到。在 $y^{(q+2)}(x)$ 连续的任何区间内，我们利用引理 5.1，并令 $s=(x-x_p)/h$，得到

$$y'(x)=\sum_{m=0}^{q}(-1)^m\binom{-s}{m}\nabla^m y'(x_p)$$

$$+(-1)^{q+1}\binom{-s}{q+1}h^{q+1}y^{(q+2)}(\xi), \quad (5-35)$$

其中 ξ 是 x, x_p 和 x_{p-q} 中的最大值和最小值之间的一点。在 x_p 到 x_{p+1} 之间积分，根据量 γ_m 的定义，我们得到

$$y(x_{p+1})-y(x_p)=h\sum_{m=0}^{q}\gamma_m\nabla^m y'(x_p)+R_q^{AB}, \quad (5-36)$$

这就是在增加余项

$$R_q^{AB}=(-1)^{q+1}h^{q+1}\int_{x_p}^{x_{p+1}}\binom{-s}{q+1}y^{(q+2)}(\xi)dx$$

后，用函数 $y'(x)$ 代入 Adams-Bashforth 公式的结果。

现在我们要用到下面两个事实：(a) $\binom{-s}{q+1}$ 在区间

$$x_p \leqslant x \leqslant x_{p+1}$$

上为定号；(b) 由引理 5.1 中的附注，$y^{(q+2)}(\xi)$ 是 x 的连续函数. 于是我们可以应用积分学中的第二中值定理[1]，结果是

$$R_q^{AB}=(-1)^{q+1}h^{q+1}y^{(q+2)}(\xi')\int_{x_p}^{x_{p+1}}\binom{-s}{q+1}dx,$$

其中 ξ' 是对应于 (x_p, x_{p+1}) 中的 x 值的一个 ξ 值，

$$x_{p-q}<\xi'<x_{p+1}.$$

由 γ_{q+1} 的定义，R_q^{AB} 便写成

$$R_q^{AB}=h^{q+2}y^{(q+2)}(\xi')\gamma_{q+1}. \tag{5-37}$$

这就是所要求的 Adams-Bashforth 公式的余项表达式. 它类似于原公式中忽略的第一项，只是现在用相应的导数表达式来代替差分.

对于 Adams-Moulton 公式，我们可以用完全类似的方法得到

$$y(x_p)-y(x_{p-1})=h\sum_{m=0}^{q}\gamma_m^*\nabla^m y'(x_p)+R_q^{AM},$$

其中

$$R_q^{AM}=h^{q+2}y^{(q+2)}(\xi)\gamma_{q+1}^*, \tag{5-38}$$

而 ξ 是满足 $x_{p-q}<\xi<x_p$ 的某个值.

如果我们想用类似的方法由公式

$$y(x_{p+1})-y(x_{p-1})=h\sum_{m=0}^{q}\kappa_m\nabla^m y'(x_p)+R_q^{NY} \tag{5-39}$$

和

$$y(x_p)-y(x_{p-2})=h\sum_{m=0}^{q}\kappa_m^*\nabla^m y'(x_p)+R_q^{MS} \tag{5-40}$$

1) 见 Taylor [1955]，118 页定理V.

来确定量 R_q^{NY} 及 R_q^{MS}，那么上面所用的方法失效，这是因为 $\binom{-s}{q+1}$ 在积分区间上现在改变了符号。我们把 R_q^{NY} 和 R_q^{MS} 的估计放到 §5.3-4 中在一般情况下来讨论．那里的引理 5.7 将给出一种普遍适用的方法．在本节中我们仅限于对特殊方法，中点法则及特殊的 Milne 公式 (5-33) 导出其误差公式．特别地，我们将证明

$$R_0^{NY} = \frac{1}{3} h^3 y'''(\xi), \quad x_{p-1} < \xi < x_{p+1} \tag{5-41}$$

和

$$R_2^{MS} = -\frac{1}{90} h^5 y^{(V)}(\xi), \quad x_{p-2} < \xi < x_p. \tag{5-42}$$

为了证明 (5-41)，不失一般性，我们可以假定 $p=0$，这样我们必须证明对 $x_{p+1} = \pm h$ 公式 (5-41) 是成立的．由于

$$R_0^{NY} = y(h) - y(-h) - 2hy'(0) \tag{5-43}$$

并利用带余项的 Taylor 展式，得到

$$y(h) = y(0) + hy'(0) + \frac{1}{2} h^2 y''(0) + \frac{1}{6} h^3 y'''(\xi_1).$$

其中 $0 < \xi_1 < 1$，类似地，

$$y(-h) = y(0) - hy'(0) + \frac{1}{2} h^2 y''(0) - \frac{1}{6} h^3 y'''(\xi_2),$$

其中 $-h < \xi_2 < 0$．代入 (5-43) 我们得到

$$R_0^{NY} = \frac{1}{6} h^3 [y'''(\xi_1) + y'''(\xi_2)].$$

由于函数 $y'''(x)$ 是连续的，从而它在区间 (ξ_2, ξ_1) 内可取 $y'''(\xi_1)$ 和 $y'''(\xi_2)$ 之间所有的值．于是，对某个 $\xi \in (\xi_2, \xi_1)$ 就有

$$y'''(\xi_1) + y'''(\xi_2) = 2y'''(\xi)$$

由此即得所要的结果。

为了证明(5-42)，不失一般性，我们可以假设 $p=1$，$x_p=h$. 于是我们就必须对某个 $\xi\in(-h,h)$ 以及

$$R_2^{MS}=y(h)-y(-h)-h\left[2y'(0)+\frac{1}{3}\nabla^2y'(h)\right]$$

证明(5-42)式成立. 容易证明 R_2^{MS} 可用定积分表达成

$$R_2^{MS}=\int_0^h\left[y'(t)+y'(-t)-2y'(0)-\frac{t^2}{h^2}\nabla^2y'(h)\right]dt.$$

现在我们考虑函数

$$F(x)=\int_0^x\left[y'(t)+y'(-t)-2y'(0)-\frac{t^2}{h^2}\nabla^2y'(h)\right]dt$$

$$-\lambda\int_0^x\left[\binom{-t/h}{4}+\binom{1+t/h}{4}\right]dt.$$

我们有 $F(0)=0$，$F(-x)=-F(x)$，并且选择 λ，使得 $F(h)=0$. 这是有可能的，因为

$$\int_0^h\left[\binom{1-t/h}{4}+\binom{1+t/h}{4}\right]dt=h\int_{-2}^0\binom{-s}{4}ds=h\kappa_4^*$$

$$=-\frac{1}{90}h\neq 0.$$

按照这样选取的 λ，由于 $F(\pm h)=F(0)=0$，$F'(\pm\xi)=0$，对某个 $\xi\in(0,h)$，我们得到

$$F(\pm h)=R_2^{MS}+\frac{1}{90}\lambda h=0. \tag{5-44}$$

但是，对于 $x=0$ 和 $x=\pm h$，

$$F'(x)=y'(x)-y'(-x)-2y'(0)-\frac{x^2}{h^2}\nabla^2y'(h)$$

$$-\lambda\left[\binom{1-x/h}{4}+\binom{1+x/h}{4}\right]$$

也为零. 于是由于 $F'(x)$ 在闭区间 $[-h,h]$ 上有五个不同的零点，其五阶导数在 $(-h,h)$ 内至少有一个零点. 我们容易

得到

$$F^{V}(x)=y^{V}(x)+y^{V}(-x)-2\lambda h^{-4},$$

因此对某个 $\xi_1\in(-h,h)$，有

$$\lambda=\frac{1}{2}h^4[y^{V}(\xi_1)+y^{V}(-\xi_1)].$$

根据 $y^{V}(x)$ 的连续性，和前面一样，对于 ξ_1 和 $-\xi_1$ 之间的某个 ξ 值，则有

$$\lambda=h^4y^{V}(\xi).$$

把这个值代入 (5-44)，便得 (5-42).

5.1-4. *建立在数值微分基础上的方法*. 建立在积分基础上的方法是通过在恒等式

$$y(x+k)-y(x)=\int_x^{x+k}f(t,y(t))dt$$

中用一个插值多项式代替函数 $f(t,y(t))$ 并对它积分而得到的，我们预先并没有理由说为什么不能用一个插值多项式并对它微分后来代替

$$y'(x)=f(x,y(x)) \tag{5-45}$$

的左端函数 $y'(x)$. 事实上建立在微分基础上的方法在实践中要比基于积分基础上的方法适合性小得多. 因为根据下面的分析就会清楚，其原因可以在误差传播的不良性质中找到.

利用插值基点为 $x_p,x_{p-1},\cdots,x_{p-q}$ 的函数 $y(x)$的插值多项式

$$P(x)=\sum_{m=0}^{q}(-1)^m\binom{-s}{m}\nabla^m y_p,\quad s=\frac{x-x_p}{h}$$

在点 $x=x_{p-r}$ 处的微分，其结果是

$$P'(x_{p-r})=\frac{1}{h}\sum_{m=0}^{q}\delta_{r,m}\nabla^m y_p, \tag{5-46}$$

其中

$$\delta_{r,m} = (-1)^m h \frac{d}{dx}\binom{-s}{m}_{x=x_{p-r}}$$

$$= (-1)^m \frac{d}{ds}\binom{-s}{m}_{s=-r}. \tag{5-47}$$

下面我们将讨论系数 $\delta_{r,m}$ 的数值及其更进一步的性质。使方程(5-46)等于 $f(x_{p-r}, y_{p-r})$，我们得到对 $y' = f(x, y)$ 的离散近似值：

$$\sum_{m=0}^{q} \delta_{r,m} \nabla^m y_p = h f_{p-r}, \tag{5-48}$$

其左边的形式是 $(\delta_{r,0} + \delta_{r,1} + \cdots + \delta_{r,m}) y_p$，$y_{p-1}$，$\cdots y_{p-q}$ 的线性组合。如果

$$\delta_{r,0} + \delta_{r,1} + \cdots + \delta_{r,m} \neq 0, \tag{5-49}$$

那么(5-48)就表示一个对于 y_p 的方程，当 $y_{p-1}, \cdots, y_{p-q}$ 的值已知时便可求解。若 $r > 0$，则这个方程为显式，而 $r = 0$ 则为隐式。后一种情形可用迭代法来求解。因此，原则上(5-48)可以象 Adams 公式那样来使用。但是，由于数值稳定性的缘故，这个公式仅在

$$r = 0, \quad q \leqslant 6,$$
$$r = 1, \quad q \leqslant 2$$

的情形下才可应用。即使在这两种情形下这个公式还是比使用相同点数相对应的 Adams-Moulton 公式的精确度要差。

尽管如此，由于公式(5-48)除了对微分方程逐次积分的用途外还有其它用处，因此我们要详细地研究它。对于系数 $\delta_{r,m}$，如果 $m > r$，利用二项式系数的定义我们求得

$$\delta_{r,m} = \frac{d}{ds}\left\{\frac{s(s+1)\cdots(s+m-1)}{1.2\cdots, m}\right\}_{s=-r}$$

$$= \frac{1}{m!} \lim_{s \to r} s(s+1)\cdots(s+r-1)(s+r+1)$$

$$\cdots(s+m-1)=(-1)^r\frac{r!(m-r-1)!}{m!}.$$

(5-50)

对于 $m\leqslant r$，直接微分并不方便. 利用一个生成函数可以建立对所有 m 都成立的关系式. 令

$$D_r(t)=\sum_{m=0}^{\infty}\delta_{r,m}t^m,$$

则有

$$\begin{aligned}D_r(t)&=\sum_{m=0}^{\infty}(-t)^m\frac{d}{ds}\binom{-s}{m}_{s=-r}\\&=\frac{d}{ds}\left(\sum_{m=0}^{\infty}\binom{-s}{m}(-t)^m\right)_{s=-r}\\&=\frac{d}{ds}(1-t)^{-s}\Big|_{s=-r}=\frac{d}{ds}e^{-s\log(1-t)}\Big|_{s=-r},\end{aligned}$$

于是

$$D_r(t)=-\log(1-t)(1-t)^r. \tag{5-51}$$

由此得到

$$\delta_{r,0}+\delta_{r,1}t+\delta_{r,2}t^2+\cdots=\left(t+\frac{1}{2}t^2+\frac{1}{3}t^3+\cdots\right)\times\left[1-\binom{r}{1}t+\binom{r}{2}t^2-\cdots+(-1)^rt^r\right].$$

通过比较系数，我们知道 $\delta_{r,0}=0$ $(r=0,1,2,\cdots)$；此外，如果 $m\leqslant r$，

$$\delta_{r,m}=\frac{1}{m}-\frac{1}{m-1}\binom{r}{1}+\frac{1}{m-2}\binom{r}{2}-\cdots+(-1)^{m-1}\binom{r}{m-1}; \tag{5-52}$$

如果 $m>r$，则有

$$\delta_{r,m} = \frac{1}{m} - \frac{1}{m-1}\binom{r}{1} + \frac{1}{m-2}\binom{r}{2} - \cdots$$

$$+ (-1)^r \frac{1}{m-r} \qquad (5\text{-}53)$$

[(5-50)和(5-53)恒等并不显然.]把恒等式

$$D_{r+1}(t) = (1-t)D_r(t), \ (1-t)^{-1}D_r(t) = D_{r-1}(t)$$

展开成 t 的幂次并比较系数后,我们又得到

$$\delta_{r+1,m} = \delta_{r,m} - \delta_{r,m-1} \quad r \geqslant 0, m \geqslant 1, \qquad (5\text{-}54)$$

$$\delta_{r-1,m} = \delta_{r,0} + \delta_{r,1} + \cdots + \delta_{r,m}, \quad r \geqslant 1, m \geqslant 0. \qquad (5\text{-}55)$$

最后我们注意到

$$\delta_{0,m} = 1/m, \quad m \geqslant 1.$$

现在容易求出表(5.9)中的数值.

表 5.9 系数 $\delta_{r,m}$

r \ m	0	1	2	3	4	5	6
0	0	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$
1	0	1	$-\frac{1}{2}$	$-\frac{1}{6}$	$-\frac{1}{12}$	$-\frac{1}{20}$	$-\frac{1}{30}$
2	0	1	$-\frac{3}{2}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{30}$	$\frac{1}{60}$
3	0	1	$-\frac{5}{2}$	$\frac{7}{6}$	$-\frac{17}{12}$	$-\frac{1}{20}$	$-\frac{1}{60}$

微分算子(5-48)的精确度是通过把该微分方程的解代入(5-48)来度量的,我们得到

$$hy'(x_{p-r}) - \sum_{m=0}^{q} \delta_{r,m} \nabla^m y(x_p) = R^D_{r,q}. \qquad (5\text{-}56)$$

根据引理 5.2,如果 $x_{p-q} \leqslant x_r \leqslant x_p$,这里 $R^D_{r,q}$ 便是

$$R^D_{r,q} = \frac{h}{(q+1)!} y^{(q+1)}(\xi) L'(x_{p-r}), \quad x_{p-q} < \xi < x_p$$

而 $L(x)=(x-x_p)(x-x_{p-1})\cdots(x-x_{p-q})$. 由于我们可以把 $L(x)$ 写成

$$L(x)=(-1)^{q+1}h^{q+1}\binom{-s}{q+1},$$

从而

$$hL'(x_{p-r})=h^{q+1}\delta_{r,q+1},$$

于是推得

$$R_{r,q}^{D}=h^{q+1}\delta_{r,q+1}y^{(q+1)}(\xi),\quad x_{p-q}<\xi<x_p; \qquad (5\text{-}57)$$

并且由它可导出 (5-48) 的误差与忽略的第一项近似地成比例，只要差分用相应的导数来代替. 应当看到 R_q^D 是与 h^{q+1} 成比例的，但是用相同点数的隐式积分公式其余项 R_q^{MA} 和 R_q^{MS} 是与 h^{q+2} 或 h^{q+3} 成比例的.

5.2. 线性多步方法的一般讨论

在 §5.1 中所讨论的方法可以看成是公式

$$\begin{aligned}&\alpha_k y_{n+k}+\alpha_{k-1}y_{n+k-1}+\cdots+\alpha_0 y_n\\&\quad=h\{\beta_k f_{n+k}+\beta_{k-1}f_{n+k-1}+\cdots+\beta_0 f_n\},\\&n=0,1,2,\cdots\end{aligned} \qquad (5\text{-}58)$$

的特殊情形，其中 k 是一个固定的整数.

$$f_m=f(x_m,y_m)\quad(m=0,1,2,\cdots),$$

以及 α_μ 和 $\beta_\mu(\mu=0,1,\cdots,k)$ 是与 n 无关的实常数. 我们总是假定 $\alpha_k\neq 0$, $|\alpha_0|+|\beta_0|>0$. 我们就称方程 (5-58) 定义了一般的线性 k 步方法. 这个方法之所以称为是线性的，这是因为 (5-58) 关于值 f_m 是线性的；但并没有假定 f 是 y 的线性函数.

为了研究在 §5.3 和 5.4 中将要考虑的累积离散误差和舍入误差，处理一般公式 (5-58) 要比分别考虑每种特殊情况更

为有效．Dahlquist [1956] 第一个采用这种方法的同时，他导出了一个在数学上经过周密考虑的舍入误差理论．同时也导出了一些新的积分公式，而这些公式用 §5.1 中所启示的方法是不可能得到的．

在本节中，我们将处理下面的问题：为了对积分一个微分方程定义一个“好”的方法，公式 (5-58) 必须具有什么性质？定理 5.5 和 5.6 将对一个“好”的方法指出两个必要条件，接着在 §5.3 中再证明这些必要条件也是充分的．

我们从线性差分方程的一些基本事实开始讨论．

5.2-1．线性差分方程．令 k 是一个正整数，I 表示相邻整数的集合，并设函数 $F_n(\eta_0,\eta_1,\cdots,\eta_k)$ 对 $n\in I$ 和 $k+1$ 个变量 $\eta_0,\eta_1,\cdots,\eta_k$ 的一切实数值都是有定义的．我们称

$$F_n(y_n,y_{n+1},\cdots,y_{n+k})=0 \qquad (5\text{-}59)$$

为 k 阶差分方程． 它表示这样的问题，即寻求一系列的数 $\{y_n\}$，使得当 $n\in I$ 时，它满足 (5-59) 式，而 $\{y_n\}$ 就称为该差分方程的解．例如

$$y_{n+1}-y_n=n,\quad -\infty<n<\infty,$$

$$n^2(y_{n+3}-y_{n+1})^2=-1,n\geqslant 27$$

都是差分方程．如果 $f(x,y)$ 对于 $x\in[a,b]$ 及所有 y 值都有定义，并设 $x_n=a+nh$，则 (5-58) 就表示一个 k 阶差分方程，而 I 由 $n\geqslant 0$，$n+k\leqslant(b-a)h^{-1}$ 所给定．

一个差分方程称为线性(和齐次)的，如果对每个 $n\in I$，函数 F_n 是变量 $\eta_0,\eta_1,\cdots,\eta_k$ 的线性(和齐次)函数． 线性差分方程可以写成

$$\alpha_{k,n}y_{n+k}+\alpha_{k-1,n}y_{n+k-1}+\cdots+\alpha_{0,n}y_n=r_{n+k},\ (5\text{-}60)$$

其中系数 $\alpha_{k,n},\cdots,\alpha_{0,n}$ 和 r_{n+k} 为所确定的 n 的已知函数，$n\in I$．即使我们允许它们是复值函数，那也不会产生额外的困难． 现在我们讨论与这种差分方程有关的一些基本性质．

其中有些性质是与 k 阶线性微分方程的理论非常类似，但也有一些显著的差别。

我们首先考虑齐次方程

$$\alpha_{k,n}y_{n+k}+\alpha_{k-1,n}y_{n+k-1}+\cdots+\alpha_{0,n}y_n=0, \quad (5-61)$$

我们假定其系数对所有 $n\geqslant 0$ 都是确定的。并规定 $\alpha_{k,n}\neq 0$，$n\geqslant 0$。显然，对于任取的初值 $y_0,y_1,\cdots,y_{k-1}$，我们都可以求得 (5-61) 的一个解，而且这个解是唯一的。特别是，对于初值 $y_0=y_1=\cdots=y_{k-1}=0$，则有对应的平凡解 $y_n=0$。如果序列 $\{y_n\}$ 和 $\{z_n\}$ 都是 (5-61) 的解，那么很明显，以 Ay_n+Bz_n 为元素组成的序列也是 (5-61) 的解，这里 A 和 B 都是任意常数。两个解 $\{y_n\}$ 和 $\{z_n\}$ 在点 $n=n_0$ 称为线性无关的，如果不存在二个不全为零的常数 A 和 B，使得

$$Ay_n+Bz_n=0, \quad n-n_0=0,1,\cdots,k-1$$

成立。更一般地，称 (5-61) 的 m 个解 $\{y_n^{(\mu)}\}(\mu=1,2,\cdots,m)$ 在 $n=n_0$ 是线性无关的，如果不存在 m 个不全为零的常数 $A_\mu(\mu=1,2,\cdots,m)$，使得关系式

$$A_1y_n^{(1)}+A_2y_n^{(2)}+\cdots+A_my_n^{(m)}=0$$

对 $n-n_0=0,1,\cdots,k-1$ 同时成立。按初等矩阵代数，这个条件就等价于矩阵

$$M=\begin{pmatrix} y_{n_0}^{(1)} & y_{n_0+1}^{(1)}\cdots y_{n_0+k-1}^{(1)} \\ y_{n_0}^{(2)} & y_{n_0+1}^{(2)}\cdots y_{n_0+k-1}^{(2)} \\ \cdots\cdots & \cdots\cdots\cdots\cdots \\ y_{n_0}^{(m)} & y_{n_0+1}^{(m)}\cdots y_{n_0+k-1}^{(m)} \end{pmatrix} \quad (5-62)$$

的秩为 m。由此可见在任一点处线性无关解的最大个数是 k。

若 (5-61) 的 k 个解组在 $n=n_0$ 是线性无关的，则我们称这 k 个解组为 (5-61) 在 $n=n_0$ 的一个基本组。这种基本组原则上是容易构造的。例如，构造这样 k 个解

$$\{y_n^{(\mu)}\} \ (\mu=1,\cdots,k),$$

使它的矩阵M变为单位矩阵就行了. 在一些特殊情形下，选取其它的基本组也许会更加方便.

如果假定

$$\alpha_{0,n} \neq 0, \quad n \geq n_0, \tag{5-63}$$

则可以证明（见问题 10）在$n = n_0$的基本组对所有$n \geq n_0$也是基本组. 另一方面，如果条件(5-63)不成立，则$n = n_0$的基本组未必是$n \geq n_0$的基本组. 但是，以下的定理，即使没有条件(5-63)也是成立的.

定理 5.1. 如果当$n \geq n_0$时，序列$\{y_n\}$是(5-61)的解，则当$n \geq n_0$时$\{y_n\}$可以表为在$n = n_0$的基本组解的线性组合.

证. 令$\{y_n^{(\mu)}\}(\mu = 1, \cdots, k)$是给定的基本组. 由条件

$$\sum_{\mu=1}^{k} A_\mu y_n^{(\mu)} = y_n, \quad n - n_0 = 0, \cdots, k - 1 \tag{5-64}$$

可确定常数$A_1, \cdots, A_k$. 因为这k个解$\{y^{(\mu)}\}$在$n = n_0$是线性无关的，所以这k个未知量A_μ的k阶线性方程组的系数行列式异于零. 因此该方程组有唯一的解$A_1, \cdots, A_k$. 如果我们规定

$$z_n = y_n - \sum_{\mu=1}^{k} A_\mu y_n^{(\mu)}, \quad n - n_0 = 0, 1, \cdots,$$

则差分方程(5-61)解的线性组合构成的序列$\{z_n\}$也是(5-61)的一个解. 由于$z_{n_0} = z_{n_0+1} = \cdots = z_{n_0+k-1} = 0$，这个解对于$n \geq n_0$恒等于零. 从而(5-64)对所有$n \geq n_0$均成立.

在定理 5.1 的基础上，我们可以说(5-61)对于$n \geq n_0$通解的形式为$\sum_{\mu=1}^{k} A_\mu y_n^{(\mu)}$，其中序列$\{y_n^{(\mu)}\}$在$n = n_0$构成一个基本组，而常数$A_\mu$是任意的. 显然对应于非齐次方程

$$\alpha_{k,n}y_{n+k}+\alpha_{k-1,n}y_{n+k-1}+\cdots+\alpha_{0,n}y_n=\gamma_{n+k} \quad (5\text{-}65)$$

的通解可以写成 $\{y_n+z_n\}$ 的形式，其中 $\{y_n\}$ 表示齐次方程（相应的 $\gamma_{n+k}=0$）的通解，$\{z_n\}$ 是非齐次方程的某个特解。非齐次方程的特解可以利用初值 $z_0=z_1=\cdots=z_k=0$，通过解 (5-65) 求得．我们可以通过齐次方程的某些解用 Duhamel 原理的离散模拟方法把这个解表示出来．对于

$$m=0,1,2,\cdots,$$

令序列 $\{y_{n,m}\}$ 满足条件

$$\begin{aligned}&y_{n,m}=0, \qquad n=0,1,\cdots,m-1,\\&y_{m,m}=1/\alpha_{k,m}\end{aligned} \quad (5\text{-}66)$$

并设对于 $n\geqslant m$，它是 (5-61) 的一个解，则有：

定理 5.2. 对于 $n>k$，满足 $z_0=z_1=\cdots=z_{k-1}=0$ 的 (5-65) 的解可以表示成

$$z_n=\sum_{m=k}^{n}\gamma_m y_{n,m}. \quad (5\text{-}67)$$

证．由 $y_{n,m}$ 的定义推出，

$$\alpha_{k,n}y_{n+k,m}+\alpha_{k-1,n}y_{n+k-1,m}+\cdots+\alpha_{0,n}y_{n,m}=\begin{cases}0, & n\neq m-k,\\1, & n=m-k.\end{cases} \quad (5\text{-}68)$$

因为对于 $m>n$，$y_{n,m}=0$，所以如果没有收敛性困难的话，(5-67) 可以换成

$$z_n=\sum_{m=k}^{\infty}\gamma_m y_{n,m}.$$

利用 (5-68)，我们就得到所要的结果

$$\alpha_{k,n}z_{n+k}+\alpha_{k-1}z_{n+k-1}+\cdots\alpha_{0,n}z_n=\sum_{m=k}^{\infty}\gamma_m\{\alpha_{k,n}y_{n+k,m}+\cdots+\alpha_{0,n}y_{n,m}\}=\gamma_{n+k}.$$

通过显式地构造常系数线性差分方程

$$\alpha_k y_{n+k}+\alpha_{k-1}y_{n+k-1}+\cdots+\alpha_0 y_n=0 \quad (5\text{-}69)$$

在 $n=0$ 的基本组来结束这一节，其中 $\alpha_k\neq 0$，$\alpha_0\neq 0$. 用类似于常系数常微分方程的方法，对于适当选取的 λ，它具有形式为 $e^{\lambda x}$ 的解.我们首先将求得 (5-69) 具有形式 $y_n=e^{\lambda n}$ 的全部解，其中 λ 是待定的. 实际上，写成 $e^{\lambda}=\zeta$ 更加方便，于是 $y_n=\zeta^n$，且容易定出 ζ. 因为我们要的是非平凡解，所以必须要求 $\zeta\neq 0$. 把 $y_n=\zeta^n$ 代入到 (5-69)，我们得到

$$\alpha_k\zeta^{n+k}+\alpha_{k-1}\zeta^{n+k-1}+\cdots+\alpha_0\zeta^n=0,$$

或者消去 ζ^n，

$$\alpha_k\zeta^k+\alpha_{k-1}\zeta^{k-1}+\cdots+\alpha_0=0.$$

由此可见，如果 $y_n=\zeta^n$ 是 (5-69) 的一个解，那么 ζ 一定是特征多项式

$$\rho(\zeta)=\alpha_k\zeta^k+\alpha_{k-1}\zeta^{k-1}+\cdots+\alpha_0$$

的一个(实或复)根. 相反，把我们的步骤倒推过去容易证明，若 ζ 满足 $\rho(\zeta)=0$，则 $y_n=\zeta^n$ 是该差分方程的一个解. 现在我们分如下两种情形.

i) 多项式 $\rho(\zeta)$ 有 k 个不同的根. 设这些根为

$$\zeta_1,\zeta_2,\cdots,\zeta_k.$$

我们已经证明了，以这些数 $y_n^{(\mu)}=\zeta_\mu^n$ 构造的 k 个序列都是 (5-69) 的解. 由于 $\alpha_0\neq 0$，故 $\zeta=0$ 不是一个根，从而 $\{y_n^{(\mu)}\}$ 没有一个是平凡解. 这些解的线性无关性是根据矩阵(5-62)对于 $n_0=0$ 化成

$$\begin{pmatrix} 1 & \zeta_1\cdots\zeta_1^{k-1} \\ 1 & \zeta_2\cdots\zeta_2^{k-1} \\ \cdots & \cdots\cdots\cdots \\ 1 & \zeta_k\cdots\zeta_k^{k-1} \end{pmatrix}$$

得到的. 该矩阵[1]行列式的值为 $\Pi_{i<j}(\zeta_i-\zeta_j)\neq 0$.

---

1) 这就是 Vandermonde 行列式. (参阅 Birkhoff 和 MacLane [1953], p.303).

ii) 多项式 $\rho(\zeta)$ 有重根． 如果多项式 $\rho(\zeta)$ 有相异的根 $\zeta_1,\zeta_2,\cdots,\zeta_m$，并设 $m<k$，那么由序列 $\{\zeta_\mu^n\}(\mu=1,\cdots,m)$ 还不足以构成基本组．其缺少的解可用常微分方程理论中熟悉的方法来求得．譬如说，如果 ζ 是二重根，那么我们可以把 ζ 看成是由两个相异的单根 ζ 和 $\zeta+\varepsilon$ 组成的，其中 ε 是一个小量． 对任何 ε，以 $\varepsilon^{-1}[(\zeta+\varepsilon)^n-\zeta^n]$ 为元素形成的序列也是一个解，这是因为它是两个解的线性组合．从而我们可以期望当 $\varepsilon\to 0$ 时得到以 $y_n=n\zeta^{n-1}$ 为元素的极限序列是对应于 $\varepsilon=0$ 的一个解．这一事实可以由下面的定理得到证实，定理包含在 (i) 下面的陈述来作为一种特殊情形．

定理 5.3. 设 $\zeta_1,\zeta_2,\cdots,\zeta_m(m\leqslant k)$ 为特征多项式 $\rho(\zeta)$ 的互异的根，又令 ζ_μ 为 p_μ 重根．则具有元素为

$$
\begin{aligned}
&y_n=\zeta_\mu^n,\\
&y_n=n\zeta_\mu^n,\\
&\cdots\\
&y_n=n(n-1)\cdots(n-p_\mu+2)\zeta_\mu^n,\quad \mu=1,2,\cdots,m
\end{aligned}
\tag{5-70}
$$

的 $k=\Sigma p_\mu$ 组序列构成差分方程 (5-69) 的解在 $n=0$ 处的一个基本组．

证． 我们首先证明，由 (5-70) 式确定的全体序列都是 (5-69) 的解． 如果 ζ_μ 是 $\rho(\zeta)$ 的 p_μ 重根，那么 ζ_μ 也是函数 $f(\zeta)=\zeta^n\rho(\zeta)$ 的 p_μ 重根，其中 n 是一个任意的非负整数．因此，$\rho(\zeta)$ 的前 $p_\mu-1$ 阶导数在 $\zeta=\zeta_\mu$ 处为零． 由此得到关系式

$$\alpha_k\zeta_\mu^{n+k}+\alpha_{k-1}\zeta_\mu^{n+k-1}+\cdots+\alpha_0\zeta_\mu^n=0,$$

$$
\begin{aligned}
\alpha_k(n+k)\zeta_\mu^{n+k-1}&+\alpha_{k-1}(n+k-1)\zeta_\mu^{n+k-2}\\
&+\cdots+\alpha_0 n\zeta_\mu^{n-1}=0,
\end{aligned}
$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\alpha_k(n+k)(n+k-1)\cdots(n+k-p_\mu+2)\zeta_\nu^{n+k-p_\mu+1}$$
$$+\cdots+\alpha_0 n(n-1)\cdots(n-p_\mu+2)\zeta_\nu^{n-p_\mu+1}=0.$$
而这些关系式就相当于说由（5-70）所确定的 p_μ 个数来求解差分方程（5-69）。

剩下来要证明由（5-69）确定的解是线性无关的．通过计算（见本章末问题 12）我们可以证明，$n=0$ 时矩阵（5-62）的行列式有非零值

$$\prod_{1\leqslant\mu<\nu\leqslant m}(\zeta_\mu-\zeta_\nu)^{p_\mu+p_\nu}\prod_{\mu=1}^{m}(p_\mu-1)!!$$

其中 $0!!=1$，$k!!=k!(k-1)!\cdots1!$，$k=1,2,\cdots$．定理 5.3 证毕。

整个这一节我们没有假设差分方程的系数为实数．如果差分方程的系数和初值条件是实的，则其解当然也是实的．但是多项式 $\rho(\zeta)$ 却仍然可以有复根．这样基本组的某些函数将以复的形式出现．类似于常微分方程的情形，我们可以得到仅含有实函数的等价的基本组．如果 ζ 是实多项式 $\rho(\zeta)$ 的一个复根，则其共轭复数 $\bar{\zeta}$ 也是一个根，且重数相同．因此，以（5-70）形成的共轭复数为元素的序列也是解，从而序列（5-70）的实部和虚部，按定义都是实的，它们本身也是差分方程的解．例如，对于差分方程 $y_{n+2}+y_n=0$，公式（5-70）给出了复基本组 $y_n=i^n$，$y_n=(-i)^n$．此时，我们可以用实的基本组 $y_n=\cos n\left(\frac{\pi}{2}\right)$，$y_n=\sin n\left(\frac{\pi}{2}\right)$ 来代替它们．

5.2-2．*近似差分方程解的存在性和唯一性*．我们研究下面的问题：假设 $f(x,y)$ 满足定理 1.1 中的条件，并设 $\alpha_k\neq0$．对任取的初值 $y_0,y_1,\cdots,y_k$，差分方程（5-58）有唯一的解 $\{y_n\}(x_n\in[a,b])$？显然，如果我们能够证明，对任何值

$$y_n,y_{n+1},\cdots,y_{n+k-1},$$

关系式(5-58)看成对 y_{n+k} 的方程有唯一的解，则回答是肯定的。若 $\beta_k=0$，这是显而易见的，因为(5-58)式可以把 y_{n+k} 显式地表示成 $y_n,\cdots,y_{n+k-1}$ 的函数，该函数对其自变量的一切值都是确定的。但是，如果 $\beta_k\neq 0$，y_{n+k} 还出现在右端项中[作为 $f_{n+k}=f(x_{n+k},y_{n+k})$ 的一个自变量]，并且(5-58)作为 y_{n+k} 的方程(一般是非线性的)，它可以有几个解或者根本没有解。如果存在唯一解，那么就产生了求这个解的实际问题。

为了分析这个问题，我们把(5-58)写成

$$y=F(y) \tag{5-71}$$

其中 $y=y_{n+k}$，

$$F(y)=h\frac{\beta_k}{\alpha_k}f(x_{n+k},y)+c, \tag{5-72}$$

$$c=\frac{1}{\alpha_k}\{h[\beta_{k-1}f_{n+k-1}+\cdots+\beta_0f_n]$$
$$-\alpha_{k-1}y_{n+k-1}-\cdots-\alpha_0y_n\}.$$

在 §5.1-2 中描述的迭代过程可以建立起一般情形，而取形式为

$$y^{(\nu+1)}=F(y^{(\nu)}),\quad \nu=0,1,2,\cdots, \tag{5-73}$$

其中 $y^{(0)}$ 是一个适合的首次近似值。下面的定理使我们不仅能够证明(对充分小的 h 值)序列 $\{y^{(\nu)}\}$ 收敛于(5-71)的解，并且这个解是唯一的。

定理 5.4. 设函数 $F(y)$ 在 $-\infty<y<\infty$ 上是确定的，并设存在常数 K，使得 $0\leqslant K<1$ 并且对任意的 y^* 和 y 都有

$$|F(y^*)-F(y)|\leqslant K|y^*-y|, \tag{5-74}$$

则下面的结论是成立的：

i) 方程(5-71)有唯一解 y;

ii) 对任意 $y^{(0)}$，由 (5-73) 确定的序列收敛于 y;

iii) 对 $\nu=1,2,\cdots$，有估计式

$$|y-y^{(\nu)}|\leqslant\frac{K}{1-K}|y^{(\nu)}-y^{(\nu-1)}|$$
$$\leqslant\frac{K^\nu}{1-K}|y^{(1)}-y^{(0)}| \tag{5-75}$$

成立。

如果 $F(y)$ 是由 (5-72) 确定的，又 $f(x,y)$ 关于 y 满足带有常数 L 的 Lipschitz 条件，那么条件 (5-74) 是满足的，取

$$K=\left|\frac{h\beta_k}{\alpha_k}\right|L. \tag{5-76}$$

而对于所有充分小的 h 值，$K<1$.

定理 5.4. 的证明．设 y 和 y^* 是 (5-71) 的两个解，则

$$y=F(y),\quad y^*=F(y^*).$$

从第二个关系式减去第一个关系式且利用 (5-74)，得到

$$|y^*-y|\leqslant K|y^*-y|.$$

由于 $|K|<1$，故要上式成立，只有 $y^*=y$. 因此 (5-71) 至多有一个解． 为了证明解的存在性，由 (5-73) 减去关系式 $y^{(\nu)}=F(y^{(\nu-1)})$ 并利用 (5-74)，便得

$$|y^{(\nu+1)}-y^{(\nu)}|\leqslant K|y^{(\nu)}-y^{(\nu-1)}|.$$

重复使用这个估计式，就有

$$|y^{(\nu+\mu)}-y^{(\nu+\mu-1)}|\leqslant K^\mu|y^{(\nu)}-y^{(\nu-1)}|$$

和

$$|y^{(\nu+1)}-y^{(\nu)}|\leqslant K^\nu|y^{(1)}-y^{(0)}|. \quad (\nu,\mu=1,2,\cdots)$$

从而对任何正整数 μ，有

$$|y^{(\nu+\mu)}-y^{(\nu)}|\leqslant|y^{(\nu+\mu)}-y^{(\nu+\mu-1)}|+\cdots+|y^{(\nu+1)}-y^{(\nu)}|$$
$$\leqslant(K^\mu+K^{\mu-1}+\cdots+K)|y^{(\nu)}-y^{(\nu-1)}|$$
$$\leqslant\frac{K^\nu}{1-K}|y^{(1)}-y^{(0)}|. \tag{5-77}$$

任給 $\varepsilon > 0$，存在一个整数 ν_0，使得

$$\frac{K^{\nu}}{1-K}|y^{(1)}-y^{(0)}|<\varepsilon$$

对一切 $\nu > \nu_0$ 成立．于是我们证明了序列 $\{y^{(\nu)}\}$ 满足关于收敛性的 Cauchy 准则．因此有一个有限的极限． 在(5-73)中令 $\nu \to \infty$，根据 $F(y)$ 满足 Lipschitz 条件因而是连续函数，我们得到

$$y=\lim_{\nu\to\infty} y^{(\nu)}=\lim_{\nu\to\infty} F(y^{(\nu)})=F(\lim_{\nu\to\infty} y^{(\nu)})=F(y).$$

于是极限 y 被认为是一个解． 并且由于已经建立了唯一性，故 y 是(5-71)的唯一解．在(5-77)中固定 ν 后，令 $\mu\to\infty$，便得估计式(5-75)．

所建立的定理实际上只是在 Banach 空间中关于函数方程的古典定理的一种十分特殊的情形(例如，见 Collatz[1960], p.38)． 在这里所考虑的情形中，这些结论都可以给出明显的几何解释．

5.2-3．*多步方法的收敛性*．设函数 $f(x,y)$ 满足存在性定理 1.1 中的条件．如果 h 充分小，并且给定了初值

$$y_0, y_1, \cdots, y_{k-1},$$

那么只要 $a \leqslant x_n \leqslant b$ 时，差分方程(5-58)唯一地确定了序列 $\{y_n\}$． 虽然所用的记号只强调依赖于离散变量 n，但该序列的元素 y_n 也与 h 有关． 其一，这是因为 h 在差分方程(5-58)中表示一个参数；其二，因为该序列的初值通常是 h 的函数：

$$y_\mu=\eta_\mu(h), \quad \mu=0,1,\cdots,k-1. \qquad (5\text{-}78)$$

对一个"好"的方法来说，在适当选择开始值以后，我们希望，当 $h\to 0$ 及 $x_n=x$ 时，由此产生的数值 y_n 趋向于所要的精确解在点 x 处的数值． 下面的定义阐述了收敛性的直观概念．

定义. 由 (5-58) 定义的线性多步法称为收敛的，如果下面的陈述对所有满足定理 1.1 中条件的一切函数 $f(x,y)$ 及一切 η 值都成立. 若 $y(x)$ 表示初值问题

$$y' = f(x,y), \quad y(a) = \eta$$

的解，则

$$\lim_{\substack{h\to 0\\ x_n = x}} y_n = y(x) \tag{5-79}$$

对所有 $x \in [a,b]$ 和一切具有满足

$$\lim_{h\to 0}\eta_\mu(h) = \eta, \quad \mu = 0,1,\cdots,k-1 \tag{5-80}$$

的初值 $y_\mu = \eta_\mu(h)$ 差分方程 (5-58) 的解 $\{y_n\}$ 都成立. 应当注意，这个定义要求条件 (5-79) 不仅对具有由精确初值——这些值当然满足 (5-80)——确定的序列 $\{y_n\}$ 是满足的，而且也对所有那些当 $h \to 0$ 时具有以精确初值 η 为极限的初值所确定的序列也是满足的. 我们提出这一个更加严格的条件，乃是因为在实际中几乎决不可能用数学上的精确值来开始计算.

5.2-4. 收敛性；稳定性条件. 在本节和下两节中，我们将考虑收敛性定义的一些推论，在这方面，把差分方程(5-58)与多项式

$$\begin{aligned}\rho(\zeta) &= \alpha_k\zeta^k + \alpha_{k-1}\zeta^{k-1} + \cdots + \alpha_0\\ \sigma(\zeta) &= \beta_k\zeta^k + \beta_{k-1}\zeta^{k-1} + \cdots + \beta_0\end{aligned} \tag{5-81}$$

联系起来是方便的.

反过来，如果给出了两个多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$，我们就可以把它们与差分方程 (5-58) 联系起来.

定理 5.5. 线性多步法 (5-58) 收敛的必要条件是有关的多项式 $\rho(\zeta)$ 的所有根的模都不超过 1，而模为 1 的根是单根.

这种加在 $\rho(\zeta)$ 上的条件就称为稳定性条件.

证． 设方法是收敛的，则它对于初值问题

$$y' = 0, \ y(0) = 0$$

是收敛的，其精确解是 $y(x) = 0$． 对于这个问题来说，(5-58) 便化成常系数差分方程

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = 0. \qquad (5-82)$$

如果这个方法是收敛的，那么根据 (5-80)，

$$\lim_{n\to\infty} y_n = 0, \quad h = x/n \qquad (5-83)$$

对任何 $x > 0$，以及满足

$$\lim_{h\to 0} \eta_\mu(h) = 0 \quad \mu = 0, 1, \cdots, k-1 \qquad (5-84)$$

的 (5-82) 的所有解 $\{y_n\}$ 都是成立的，其中 $y_\mu = \eta_\mu(h)$． 令 $\zeta = re^{i\varphi}$ ($r \geqslant 0$, $0 \leqslant \varphi < 2\pi$) 为 $\rho(\zeta)$ 的一个根，则由定理 5.3，数

$$y_n = hr^n \cos n\varphi \qquad (5-85)$$

确定 (5-82) 的一个解，并且满足 (5-84)． 如果方法收敛的话，那么 (5-83) 式必须成立． 若 $\varphi = 0$ 或 $\varphi = \pi$，则立即导出 $r \leqslant 1$． 若 $\varphi \neq 0$，$\varphi \neq \pi$，我们注意到

$$\frac{y_n^2 - y_{n+1} y_{n-1}}{\sin^2 \varphi} = h^2 r^{2n},$$

由于当 $n \to \infty$，$h = x/n$ 时，它的左端项趋于零，故其右端项也必须趋于零，这样又得到 $r \leqslant 1$． 这就证明了定理 5.5 中结论的第一部分． 为了证明它的第二部分，设 $\zeta = re^{i\varphi}$ 为 $\rho(\zeta)$ 的重数大于 1 的根，则再根据定理 5.3，数

$$y_n = h^{1/2} n r^n \cos n\varphi \qquad (5-86)$$

为 (5-82) 的一个解． 因为 $|\eta_\mu(h)| = |y_\mu| \leqslant h^{1/2} \mu r^\mu$ ($\mu = 0, 1, \cdots, k-1$)，所以它们也满足 (5-84)． 于是对于一个收敛的方法，它们必须满足 (5-84)． 如果 $\varphi = 0$ 或 $\varphi = \pi$，那么对 $h = x/n$，我们就有 $|y_n| = x^{1/2} n^{1/2} r^n$，由此即得 $r < 1$． 若

$\varphi \neq 0$，$\varphi \neq \pi$，则我们可以利用关系式

$$\frac{z_n^2 - z_{n+1}z_{n-1}}{\sin^2\varphi} = r^{2n},$$

其中 $z_n = n^{-1}h^{-1/2}y_n$. 由(5-83)，当 $n \to \infty$ 时，$z_n \to 0$，故该关系式的左端项趋于零，并且得到 $r < 1$. 于是定理得证.

也许有人会认为，对于 $r > 1$ 及 $r \geq 1$，由(5-85)和(5-86)所定义的序列，其发散性与这种开始值的选取有关. 其实不然. 如果多项式 $\rho(\zeta)$ 违反了稳定性条件，那么可以证明，在某种意义下，即使(5-84)是成立的，"几乎所有"差分方程(5-82)的解却是发散的.

作为一个不稳定方法(即违反稳定性条件的方法)的数值例子，我们采用选择 $q = 2$，$r = 2$ 的微分方法(5-48)来考察初值问题 $y' = y$，$y(0) = 1$ 的解. 我们注意到差分方程(5-48)中的项 $\nabla^m y_{n+k}$ 对应于多项式 $\rho(\zeta)$ 的项 $\zeta^k(1-\zeta^{-1})^m$. 于是对于问题中的这个方法，求得

$$\begin{aligned}\rho(\zeta) &= \zeta^2[\delta_{21}(1-\zeta^{-1}) + \delta_{22}(1-\zeta^{-1})^2] \\ &= \zeta(\zeta-1) - \frac{3}{2}(\zeta-1)^2 \\ &= -\frac{1}{2}\zeta^2 + 2\zeta - \frac{3}{2}\end{aligned}$$

的根为 $\zeta = 1$，$\zeta = 3$. 方程(5-58)为

$$-\frac{1}{2}y_{n+2} + 2y_{n+1} - \frac{3}{2}y_n = hy_n,$$

用 $h = 0.1$ 和开始值 $y_0 = 1.00000$，$y_1 = 1.10517$ (与给定位数的精确解一致)求这个差分方程的数值解，我们得到在表5.10中给出的数值. 表中数值很快发散是明显的，甚至对更小的 h 值这种发散现象会发生得更快.

至于在§5.1中讨论过的其它一些方法，我们看到，建立

表 5.10

x_n	y_n	$y_n - e^{x_n}$
0	1.00000	0.00000
0.1	1.10517	0.00000
0.2	1.22068	−0.00072
0.3	1.34618	−0.00368
0.4	1.47854	−0.01329
0.5	1.60638	−0.04234
0.6	1.69419	−0.12793
0.7	1.63634	−0.37741
0.8	1.12395	−1.10159
0.9	−0.74049	−3.20009
1.0	−6.55860	−9.27688

在积分基础上的所有方法都满足稳定性条件。这是由于这些方法的多项式 $\rho(\zeta)$ 形式为 $\zeta^k - \zeta^{k-p}$ $(1 \leqslant p \leqslant k)$，因此在单位圆上有 p 个互异的根，而 $\zeta = 0$ 为 $k - p$ 重根。建立在数值微分基础上的公式却没有这种简单的一般结果，因为 $\rho(\zeta)$ 的根的位置并不是那么容易确定的。通过数值计算或应用 Hurwitz 提出的代数准则，便可以证明当 $r = 0$ 时（在最前面的点处的微分）这些方法对于 $q = 1, 2, \cdots, 6$ 是稳定的。对于 $q = 7$ 以及所有更大的 q，这些方法非常可能是不稳定的。当 $r = 1$ 时，稳定性在 $q = 3$ 时失效，$r = 2$ 时正如上面所看到的那样，在 $q = 2$ 时失效。

5.2-5. *相伴差分算子；阶和误差常数.* 稳定性条件具有这样的意义，即防止小的初始误差在运算中迅速增长，以致收敛性受到破坏。显然，仅仅有稳定性并不能够保证收敛性。因此必须加上其它条件才能保证差分方程 (5-58) 为微分方程 $y' = f(x, y)$ 的一个好的近似。

在考虑度量单步方法精确度的时候，我们曾考察表达式

$$y(x+h)-y(x)-h\Phi(x,y(x),h),$$

其中 $y(x)$ 为给定微分方程的一个解． 这个量作为 h 的函数来说，如果越小的话，那么方法的精确度就越高．同样地，如果 (5-58) 要定义一个好方法，那么我们期望，如果 h 较小并用 $y(x_m)$ 代替 y_m 的话，(5-58) 式两边的偏差要小，这里 $y(x)$ 是给定微分方程的一个精确解．我们把差分算子

$$\begin{aligned}L[y(x);h]=\ &\alpha_k y(x+kh)+\alpha_{k-1}y(x+(k-1)h)+\cdots\\&+\alpha_0 y(x)-h\{\beta_k y'(x+kh)+\beta_{k-1}y'(x+(k-1)h)\\&+\cdots+\beta_0 y'(x)\}\end{aligned}\tag{5-87}$$

和 (5-58) 式联系起来． 它可以看成是"作用"于任何可微函数 $y(x)$ 上的一个线性算子．而暂时，我们只把算子 L 作用于具有充分高阶连续导数的函数．这样我们就可以按 h 的幂次来展开 $L[y(x);h]$，并且展开式可以取任意多项．由于

$$y(x+mh)=y(x)+mhy'(x)+\frac{1}{2}m^2h^2y''(x)+\cdots,$$

$$hy'(x+mh)=hy'(x)+mh^2y''(x)+\cdots,$$

得到

$$\begin{aligned}L[y(x);h]=C_0y(x)+C_1hy'(x)+\cdots\\+C_qh^qy^{(q)}(x)+\cdots,\end{aligned}\tag{5-88}$$

其中系数 $C_q(q=0,1,\cdots)$ 是常数与所选择的函数 $y(x)$ 无关：

$$C_0=\alpha_0+\alpha_1+\alpha_2+\cdots+\alpha_k,$$

$$C_1=\alpha_1+2\alpha_2+\cdots+k\alpha_k-(\beta_0+\beta_1+\beta_2+\cdots+\beta_k),$$

$$\begin{aligned}C_q=\ &\frac{1}{q!}(\alpha_1+2^q\alpha_2+\cdots+k^q\alpha_k)\\&-\frac{1}{(q-1)!}(\beta_1+2^{q-1}\beta_2+\cdots+k^{q-1}\beta_k),\end{aligned}$$

$q=2,3,\cdots.$

若 $C_0 = C_1 = \cdots = C_p = 0$，而 $C_{p+1} \neq 0$，则称给定的差分算子 (5-87) 为 p 阶的．与单步法的情形一样，这个阶可以看成方法精确度的一种粗糙度量．在单步方法的定义和多步方法的定义的少数情形中（如第 2 章问题 3 中所考虑的 Euler方法和梯形法则），易见，所给出的这两个阶的定义是一致的．

实际上，差分方程 (5-58) 与方程

$$\alpha_k y_{n+t+k} + \cdots + \alpha_0 y_{n+t} = h(\beta_k f_{n+t+k} + \cdots + \beta_0 f_{n+t}) \tag{5-89}$$

是完全等价的，其中 t 是任意固定的(正的或负的)整数．按上面进行，我们可以把差分算子

$$\begin{aligned} L_t[y(x);h] = {} & \alpha_k y(x+(t+k)h) + \cdots + \alpha_0 y(x+th) \\ & - h\{\beta_k y'(x+(t+k)h) + \cdots + \beta_0 y'(x+th)\} \end{aligned} \tag{5-90}$$

与 (5-89) 式联系起来，并把它的阶数定义为它以 h 为幂次的 Taylor 展式中第一个非零项的幂次减 1．一个重要的事实是，阶数 p 和常数 C_{p+1} 并不依赖于 t．因为以 h 为幂次的展开式 (5-90) 和展开式 (5-87) 是等价的，只是把其中的 $y(x)$ 换成了 $y(x+th)$．如果 L 是 p 阶的，那么求得

$$\begin{aligned} L_t[y(x);h] = L[y(x+th);h] = {} & C_{p+1}h^{p+1}y^{(p+1)}(x+th) \\ & + O(h^{p+2}). \end{aligned}$$

由于假定了 $y(x)$ 是充分可微的，故

$$y^{(p+1)}(x+th) = y^{(p+1)}(x) + O(h).$$

从而

$$L_t[y(x);h] = C_{p+1}h^{p+1}y^{(p+1)}(x) + O(h^{p+2}).$$

这就证明了这个论断．甚至在证明中并没有用到 t 是一个整数的假定． 同时这个论证表明，常数 $C_{p+2}, C_{p+3}, \cdots$ 一般地必须是依赖于 t 的．

作为一个例子，考察中点法则[在(5-25)式中，令 $q=0$]，按照一般形式 (5-58)，它可以写成：

$$y_{n+2}-y_n=h\cdot 2f_{n+1},$$

相应的算子 (5-87) 为

$$\begin{aligned}L[y(x);h]&=y(x+2h)-y(x)-2hy'(x+h)\\&=2hy'+\frac{(2h)^2}{2}y''+\frac{(2h)^3}{6}y'''+\frac{(2h)^4}{24}y^{\mathrm{IV}}+\cdots\\&\quad-2hy'-2h^2y''-2\frac{h^3}{2}y'''-2\frac{h^4}{6}y^{\mathrm{iv}}-\cdots.\end{aligned}$$

我们容易得到 $C_0=C_1=C_2=0$，$C_3=\frac{1}{3}$，$C_4=\frac{1}{3}$，于是 $p=2$. 另一方面，令 $t=-1$，我们可以考虑算子

$$\begin{aligned}L_{-1}[y(x);h]&=y(x+h)-y(x-h)-2hy'(x)\\&=2hy'+2\frac{h^3}{6}y'''+2\frac{h^5}{120}y^{\mathrm{V}}+\cdots-2hy',\end{aligned}$$

又得到 $C_0=C_1=C_2=0$，$C_3=\frac{1}{3}$，因此 $p=2$，但 $C_4=0$. 从实际计算的观点出发，显然用第二种方法计算 p 和 C_{p+1} 较好，因为它在 Taylor 展式中只出现奇次项. 这个注释通常适用于 $\alpha_\mu=-\alpha_{k-\mu}$，$\beta_\mu=\beta_{k-\mu}$ 的差分方程；见本章末问题 32.

实际上，一般并不需要通过计算所有的常数 C_0，C_1，$\cdots$ 来确定 p 和 C_{p+1}. 例如，对 §5.1 中所考虑的大部分公式，我们证明了适当地确定一个 t，就有

$$L_t[y(x);h]=Kh^{q+1}y^{(q+1)}(\xi),\qquad(5\text{-}91)$$

其中 q 和 $K\neq 0$ 都是与 $y(x)$ 无关的常数，以及 ξ 是指定区间上的一点. 只要 (5-91) 成立，我们就可以推出

$$p=q,\quad C_{p+1}=K.$$

因为根据上面的说明我们有

$$L_t[y(x);h]=C_{p+1}h^{q+1}y^{(p+1)}(x)+\cdots,\qquad(5\text{-}92)$$

其中点…表示 $r > p+1$ 的那些项 $h^r y^{(r)}(x)$. 在(5-91)和(5-92)两式取 $y(x) = x^{p+1}/(p+1)!$, 这样左边的算子是恒等的. 因此 $Kh^{q+1} = C_{p+1}h^{p+1}$,因为它对所有的 h 值都必须成立,从而推出 $q = p$, $K = C_{p+1}$.

引进差分算子的阶 p 作为算子精确度的初步而又粗糙的度量. 在给定阶数的所有差分算子类中是否可以把 C_{p+1} 的值看成是精确度的一种较好的度量呢? 回答显然是否定的,因为我们可以用一个适当的常数乘以(5-58)式使 C_{p+1} 的值任意小,而这一过程并不改变由这个公式所确定的值 y_n,从而也就没有提高该方法的精确度. 但是,在公式(5-58)的这样一些改变下,我们给出的常数

$$C = \frac{C_{p+1}}{\beta_0 + \beta_1 + \cdots + \beta_k} \tag{5-93}$$

却是精确度的一种好的而且在许多方面是方便的度量. 这个常数便称为由(5-58)式所确定的方法的误差常数. 对于所有收敛的方法均可确定这个误差常数. 因为在 §5.2-6 中将

表 5.11

特殊差分算子的阶和误差常数. [q = 所用的最高阶差分阶, k = 相伴差分方程(5-58)的阶]

方　法	相伴差分算子的阶	误差常数
Adams-Bashforth	$q+1=k$	γ_{q+1}
Adams-Moulton	$q+1=k+1$	γ^*_{q+1}
Nyström $q \neq 0$	$q+1=k$	$\frac{1}{2}\kappa_{q+1}$
Nyström $q=2$	2	$\frac{1}{6}$
Milne-Simpson $q \neq 2$	$q+1=k+1$	$\frac{1}{2}\kappa^*_{q+1}$
Milne-Simpson $q=2$	4	$-\frac{1}{180}$
在 x_{n+k-r} 处的微分	$q=k$	$\delta_{r,q+1}$

指出，所有这些方法都满足条件 $\beta_0+\beta_1+\cdots+\beta_k\neq 0$.

表 5.11 中列出了 §5.1 中所讨论的一些特殊方法的阶和误差常数.

5.2-6. 收敛性；相容性条件. 在 §5.2-5 开始所提示的条件，现在叙述如下：

定理 5.6. 由 (5-58) 所确定的线性多步法收敛的一个必要条件是相伴差分算子的阶至少为 1.

阶 $p\geqslant 1$ 的条件称为相容性条件. 按 §5.2-5 中引进的常数 C_n，这个条件就相当于 $C_0=0$，$C_1=0$；相容性条件可以用多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 表达成关系式

$$\rho(1)=0,\ \rho'(1)=\sigma(1). \tag{5-94}$$

容易证明，§5.1 中所讨论的特殊方法都是相容的.

定理 5.6 的证明. 我们先来证明 $C_0=0$. 若方法是收敛的，则它对具有精确解的初值问题 $y'=0$，$y(0)=1$ 也是收敛的. 差分方程 (5-58) 又可以化成

$$\alpha_k y_{n+k}+\alpha_{k-1}y_{n+k-1}+\cdots+\alpha_0 y_n=0. \tag{5-95}$$

假定这个方法是收敛的，则取精确开始值为

$$y_\mu=1\quad(\mu=0,1,\cdots,k-1)$$

的公式 (5-95) 的解 $\{y_n\}$ 必须满足 $y_n=1$，当 $h\to 0$，$nh=x$ 时. 因为在这种情形下，y_n 不依赖于 h，这也就是说，当 $n\to\infty$ 时，$y_n\to 1$. 在 (5-95) 中，令 $n\to\infty$，我们得到

$$\alpha_k+\alpha_{k-1}+\cdots+\alpha_0=0.$$

这就等价于 $C_0=0$. 于是得到 $p\geqslant 0$.

为了证明 $C_1=0$，考察初值问题 $y'=1$，$y(0)=0$. 其精确解为 $y=x$. 差分方程 (5-58) 现在可以写成

$$\alpha_k y_{n+k}+\alpha_{k-1}y_{n+k-1}+\cdots+\alpha_0 y_n$$
$$=h(\beta_k+\beta_{k-1}+\cdots+\beta_0). \tag{5-96}$$

对于一个收敛的方法，(5-96) 的每一个解都满足

$$\lim_{h\to 0}\eta_\mu(h)=0,\quad \mu=0,1,\cdots,k-1,\tag{5-97}$$

其中 $y_\mu=\eta_\mu(h)$，同时还必须满足

$$\lim_{h\to 0}y_n=x,\quad x_n=x.\tag{5-98}$$

此外，对于一个收敛的方法，由定理 5.5 我们可以假定

$$k\alpha_k+(k-1)\alpha_{k-1}+\cdots+\alpha_1=\rho'(1)\neq 0.$$

设序列 $\{y_n\}$ 是由 $y_n=nhK$ 所确定的，其中

$$K=\frac{\beta_k+\beta_{k-1}+\cdots+\beta_0}{k\alpha_k+(k-1)\alpha_{k-1}+\cdots+\alpha_1}.$$

这个序列显然满足 (5-97) 并且容易证明它是 (5-98) 的一个解．由 $\lim nhK=xK$，$nh=x$，我们断定 $K=1$．这就相当于 $C_1=0$．从而完成了定理 5.6 的证明．

这就证明了，从所考察的例子导出的稳定性和相容性条件对于收敛性也是必要条件．鉴于这个事实，令人更加意外的是这些条件对收敛性不仅是必要的，而且也是充分的．而且，对于充分精确的开始值，一个 p 阶方法的累积离散误差，和单步法的情形一样，是 $O(h^p)$．这些结果的证明并非显然，将在 §5.3 中给出．

5.2-7．**最大阶算子的构造**．本节我们将研究下面的问题：给定一个满足 $\rho(1)=0$ 的 k 次多项式，通过选择适当的多项式 $\sigma(\zeta)$，其相伴算子 L 能达到多少阶？ 定理 5.7 给出了回答．这个答复使我们能够以更加一般的观点来考察 5.1 中所讨论的那些特殊算子．为了讨论这个问题，需要用到一些复变理论．

对形式 (5-87) 的任何差分算子，我们考虑复变量 ζ 的函数：

$$\varphi(\zeta)=(\log\zeta)^{-1}\rho(\zeta)-\sigma(\zeta).\tag{5-99}$$

函数 $\log\zeta$ 是通过将复平面沿着负实轴剖开并令 $\log 1=0$ 所

作出的单值函数，函数 $(\log\zeta)^{-1}$ 在 $\zeta=1$ 处有一个一级极点． 若算子是相容的，则这个极点就与 $\rho(\zeta)$ 的一个零点相消，函数 $\varphi(\zeta)$ 在 $\zeta=1$ 处是解析的． 更一般地，有下面的引理成立．

引理 5.3． 与多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 有关的差分算子(5.87)为 p 阶精确当且仅当函数 $\varphi(\zeta)$ 在 $\zeta=1$ 处有一个 p 级零点．

证． 设这个算子是 p 阶的，则对充分可微的函数 $y(x)$，表达式(5.87)为 $O(h^{p+1})$． 特别是，选取 $y(x)=e^x$，当 $h\to 0$ 时，就有

$$L[e^x;h]=e^x\{\rho(e^h)-h\sigma(e^h)\}=e^xC_{p+1}h^{p+1}+O(h^{p+2}), \tag{5-100}$$

其中 $C_{p+1}\neq 0$． 这就意味着在 $h=0$ 处解析函数

$$f(h)=\rho(e^h)-h\sigma(e^h)$$

有 $p+1$ 级零点，并且 $h^{-1}f(h)$ 在 $h=0$ 处有 p 级零点．由于变换 $\zeta=e^h$ 把 $h=0$ 的一个邻域一一对应地映射到 $\zeta=1$ 的一个邻域．根据复变理论中[1]的古典定理便得到函数

$$\varphi(\zeta)=(\log\zeta)^{-1}f(\log\zeta)$$

在 $\zeta=1$ 处有一个 p 级零点．

相反地，设 $\varphi(\zeta)$ 在 $\zeta=1$ 有一个 p 级零点． 则如上所述，推出函数 $f(h)=h\varphi(e^h)$ 在 $h=0$ 处有一个 $p+1$ 级零点． 因此(5-100)对某个非零常数 C_{p+1} 的值是成立的． 从而当以 $y(x)=e^x$ 代入时，$L[y(x);h]$ 的阶为 p． 由于阶只依赖于系数 α_μ 和 β_μ，从而 L 的阶为 p．

现在立得下述肯定的结果．

定理 5.7． 令 $\rho(\zeta)$ 为满足 $\rho(1)=0$ 的 k 次多项式，并

---

1) 见 Ahlfors [1953]，p.107，定理 11．

令 k' 是一个整数，$0 \leqslant k' \leqslant k$，则存在唯一的次数 $\leqslant k'$ 的多项式 $\sigma(\zeta)$，使得相伴差分算子的阶至少为 $k'+1$ 阶.

证． 函数 $(\log\zeta)^{-1}\rho(\zeta)$ 在 $\zeta=1$ 处是解析的，因此能够按 $\zeta-1$ 的非负幂次展开：

$$\frac{\rho(\zeta)}{\log\zeta}=c_0+c_1(\zeta-1)+c_2(\zeta-1)^2+\cdots. \quad (5\text{-}101)$$

如果我们令

$$\sigma(\zeta)=c_0+c_1(\zeta-1)+\cdots+c_{k'}(\zeta-1)^{k'}, \quad (5\text{-}102)$$

则函数 $\varphi(\zeta)$ 在 $\zeta=1$ 处有 $k'+1$ 重零点． 并且按引理 5.3 相伴算子的阶 $\geqslant(k'+1)$（若 $c_{k'+1}=0$，则阶超过 $k'+1$）．相反，若 L 为 $k'+1$ 阶，则 $\varphi(\zeta)$ 在 $\zeta=1$ 处有 $k'+1$ 重零点． 由 Taylor 展式的唯一性，所以 $\sigma(\zeta)$ 必须与 (5-102) 式恒等． 于是定理5.7证毕．只有 $k'=k-1$ 和 $k'=k$ 的情形是有实际意义的；前者导出了最佳显式算子，而后者导出了最佳的隐式算子．

证明的方法也提供了一个确定误差常数的方法． 若与 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 有关的差分算子为 p 阶，则由 (5-100)，

$$\rho(e^h)-h\sigma(e^h)=c_{p+1}h^{p+1}+O(h^{p+2}), \quad (5\text{-}103)$$

其中 $C_{p+1}\neq 0$． 因为当 $\zeta\to 1$ 时，

$$\log\zeta=\zeta-1+O((\zeta-1)^2),$$

从而

$$\frac{\rho(\zeta)}{\log\zeta}-\sigma(\zeta)=C_{p+1}(\zeta-1)^p+O((\zeta-1)^{p+1}). \quad (5\text{-}104)$$

因此，C_{p+1} 就等于以 $\zeta-1$ 为幂次的 $\rho(\zeta)/\log\zeta$ 的展开式中未被 $\sigma(\zeta)$ 吸收进去的第一个非零项的系数． 由 $\sigma(1)=c_0$，我们得到误差常数

$$C = c_p / c_0. \tag{5-105}$$

作为一个例子，令 $\rho(\zeta) = (\zeta - 1)(\zeta - \lambda)$，其中 λ 是实的．这是与相容性相配合最一般的二次多项式．对于

$$-1 \leqslant \lambda < 1,$$

它满足稳定性条件．我们便有

$$\frac{\rho(\zeta)}{\log \zeta} = \frac{(\zeta - 1)[1 - \lambda + (\zeta - 1)]}{\log[1 + (\zeta - 1)]}$$

$$= 1 - \lambda + \frac{3 - \lambda}{2}(\zeta - 1) + \frac{5 + \lambda}{12}(\zeta - 1)^2$$

$$- \frac{1 + \lambda}{24}(\zeta - 1)^3 + O((\zeta - 1)^4).$$

于是与上述 $\rho(\zeta)$ 有关的最佳隐式方法为

$$\sigma(\zeta) = 1 - \lambda + \frac{3 - \lambda}{2}(\zeta - 1) + \frac{5 + \lambda}{12}(\zeta - 1)^2$$

$$= -\frac{1 + 5\lambda}{12} + \frac{2 - 2\lambda}{3}\zeta + \frac{5 + \lambda}{12}\zeta^2,$$

而相应的差分方程 (5.58) 为

$$y_{n+2} - (1 + \lambda)y_{n+1} + \lambda y_n$$

$$= h\left\{\frac{5 + \lambda}{12}f_{n+2} + \frac{2 - 2\lambda}{3}f_{n+1} - \frac{1 + 5\lambda}{12}f_n\right\}. \tag{5-106}$$

对于 $\lambda \neq -1$，该方法的阶为 3，误差常数为

$$C = -\frac{1}{24}\frac{1 + \lambda}{1 - \lambda};$$

对于 $\lambda = -1$，由于该方法与 Milne 方法相同，故方法的阶数是 4．

如果我们希望用向后差分来表示 (5-58) 右端的表达式，则引进新的变量 $t = 1 - \zeta^{-1}$ 是很方便的． 为了从给定的 $\rho(\zeta)$ 导出阶 $p \geqslant k + 1$ 的隐式方法，我们令

$$\rho(\zeta) = \zeta^k R(t), \quad \sigma(\zeta) = \zeta^k S(t), \tag{5-107}$$

其中 R 和 S 都是 t 的 k 次多项式，$R(0)=0$. 考虑到

$\log\zeta=-\log(1-t)$, $\zeta-1=t(1-t)^{-1}=t+O(t^2)$,

从 (5-104) 得到

$$-\frac{R(t)}{\log(1-t)}-S(t)=C_{p+1}t^p+O(t^{p+1}). \quad (5\text{-}108)$$

于是多项式 $S(t)$ 恒等于函数 $-[\log(1-t)]^{-1}R(t)$ 在 $t=0$ 处的 k 阶 Taylor 多项式，并且 $C_{p+1}t^p$ 就等于该函数的 Taylor 展式中下一个非零项.

作为一个例子，令 $\rho(\zeta)=\zeta^k-\zeta^{k-1}$, 我们求得 $R(t)=t$, 并且由展式

$$-\frac{1}{\log(1-t)}=\gamma_0^*+\gamma_1^*t+\cdots$$

(见 §5.1-2)，我们把多项式 $S(t)$ 的系数与出现在 Adams-Moulton 公式的系数等同起来. 由于 $S(t)$ 中的项 $\zeta^k t^\mu$ 提供了 (5-58) 的右端的项 $\nabla^\mu f_{n+k}$, 因此这样产生的方法的确是与 Adams-Moulton 方法恒等的. 除了从一般的观点重新发现这种方法以外，我们证明了这是由多项式 $\rho(\zeta)=\zeta^k-\zeta^{k-1}$ 所能得到的最好的方法. 同样，可以证明 Milne-Simpson 方法是由多项式 $\rho(\zeta)=\zeta^k-\zeta^{k-2}$ 所能得到的最好的方法.

如果要得到一个 k 阶的显式方法，一个更为方便的方法是令

$$\rho(\zeta)=\zeta^k R(t)=\zeta^{k-1}\frac{R(t)}{1-t}, \sigma(\zeta)=\zeta^{k-1}S(t), \quad (5\text{-}109)$$

这里 $S(t)$ 是 $k-1$ 次的. 这样我们得到

$$-\frac{R(t)}{(1-t)\log(1-t)}-S(t)=C_{p+1}t^p+O(t^{p+1}), \quad (5\text{-}110)$$

且 $S(t)$ 可以和在 $t=0$ 的解析函数的 Taylor 多项式恒等. 令

$\rho(\zeta)=\zeta^k-\zeta^{k-1}$, $\rho(\zeta)=\zeta^k-\zeta^{k-2}$，我们又重新得到 Adams-Bashforth 和 Nyström 方法，同时证明了这些是上述问题中涉及的多项式 $\rho(\zeta)$ 所能得到的最好的显式方法。

5.2-8. 稳定算子的最大阶. 在前节中我们已经指出对给定的任意满足 $\rho(1)=0$ 的 k 次多项式 $\rho(\zeta)$，人们总是可以求得一个多项式 $\sigma(\zeta)$ 使得相伴算子至少为 $k+1$ 阶. 于是我们就得到了与给定的 $\rho(\zeta)$ 有关的差分算子的最大阶的一个下界. 现在提出这样的问题，即通过适当选择 $\rho(\zeta)$ 其差分算子所能达到的阶究竟比 $k+1$ 增大多少.

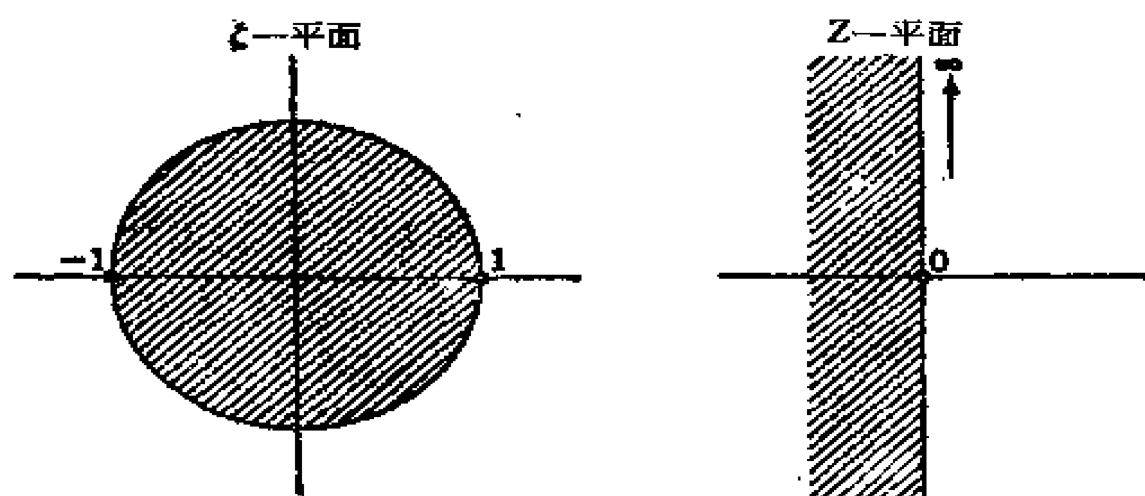


图 5.2

根据我们安排的参数个数(系数 α_μ 和 β_μ 各 $k+1$ 个)以及所要满足的条件个数暗示 p 能达到的最大值为 $2k$. 这个推测的正确性已由 Dahlquist [1956]证明. 但是，这个结果并没有什么大的实际意义. 因为现在我们将证明也是由 Dahlquist 证明的重要结果，即任何阶 $p>k+2$ 的算子必定是不稳定的. 因此满足稳定性条件的 $\rho(\zeta)$ 所能达到的最大阶为 $k+2$，而且即使如此也只是对于那些受到相当限制的一类多项式 $\rho(\zeta)$ 才有可能.

为了证明这个结果，令 $\rho(\zeta)$ 是适合相容性和满足稳定性条件的多项式[1]。为了分析地处理这些假设，我们通过线性分式变换

$$z=\frac{\zeta-1}{\zeta+1},\ \zeta=\frac{1+z}{1-z} \tag{5-111}$$

而引进一个新的变量 $z=x+iy$。这个变换把圆 $|\zeta|<1$ 映射到 $\mathrm{Re}\, z<0$ 的半平面上，把 $|\zeta|=1$ 映射到虚轴上，点 $\zeta=1$ 映射到 $z=0$，以及 $\zeta=-1$ 映射到 $z=\infty$（见图 5.2）。

现在我们考虑用以下函数来代替多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$

$$r(z)=\left(\frac{1-z}{2}\right)^k\rho\left(\frac{1+z}{1-z}\right),\ S(z)=\left(\frac{1-z}{2}\right)^k\sigma\left(\frac{1+z}{1-z}\right), \tag{5-112}$$

它们是次数 $\leqslant k$ 的多项式。若 $\rho(\zeta)$ 在点 $\zeta=\zeta_0$ 处有 p 重根，则 $r(x)$ 在点 $z=(\zeta_0-1)/(\zeta_0+1)$ 有相同重数的根，只是 $\zeta_0=-1$ 除外，而在这种情形下，$r(z)$ 的次数化成 $k-p$ 次。由于 $\zeta=1$ 是 $\rho(\zeta)$ 的一个单根，故 $z=0$ 为 $r(z)$ 的一个单根。于是

$$r(z)=a_1z+a_2z^2+\cdots+a_{k-1}z^{k-1}+a_kz^k, \tag{5-113}$$

其中 $a_1\neq 0$。$r(z)$ 的系数都是实数。不失一般性，我们可以假定

$$a_1>0, \tag{5-114}$$

因为这总可以通过一个适当的因子乘以 $\rho(\zeta)$ 来实现。在这个假设下，$r(z)$ 的其它系数便满足

$$a_\mu\geqslant 0,\quad \mu=1,\cdots,k. \tag{5-115}$$

这可以通过考察多项式 $r(x)$ 的乘积展开式而得到证明。用 $x_\nu+iy_\nu$ 表示 $r(z)$ 的根，而用 a_l 表示具有最大下标的那

1) 真正用到的仅是下面较弱的条件：$\rho(\zeta)$ 没有模大于 1 的根，而 $\zeta=1$ 的根为单根。

个非零系数，则

$$r(z)=a_l z\prod_{\lambda}(z-x_\lambda)\prod_{\mu}((z-x_\mu)^2+y_\mu^2),$$

(5-116)

其中 λ 在实根中变化，而 μ 在一对共轭复根中变化. 由稳定性，对所有根，$x_\mu\leqslant 0$. 于是由展开 (5-116) 式便可知，z 的各幂次均是乘以非负常数. 由此可见，$r(z)$ 的一切非零系数与 a_l 同号. 既然 $a_1>0$，从而它们也都是正的.

现在我们考察函数

$$p(z)=\left(\frac{1-z}{2}\right)^k\varphi\left(\frac{1+z}{1-z}\right)$$

$$=\frac{1}{\log\frac{1+z}{1-z}}r(z)-S(z).\qquad(5\text{-}117)$$

函数 $p(z)$ 在 $z=0$ 处有 p 级零点当且仅当 $\varphi(\zeta)$ 在 $\zeta=1$ 处有 p 级零点，于是按引理 5.3，这仅当与 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 有关的算子阶为 p. 因此，若算子为 p 阶，则

$$s(z)=b_0+b_1z+b_2z^2+\cdots+b_{p-1}z^{p-1},\qquad(5\text{-}118)$$

其中

$$\frac{z}{\log\frac{1+z}{1-z}}\frac{r(z)}{z}=b_0+b_1z+b_2z^2+\cdots.\qquad(5\text{-}119)$$

由于 $s(z)$ 的次数一定不大于 k，所以对 $p>k+1$ 的稳定算子的存在性要视 $b_{k+1}=\cdots=b_{p-1}=0$ 的可能性而定. 为了研究这个可能性，我们定出这些系数 b_ν. 令

$$\frac{z}{\log\frac{1+z}{1-z}}=c_0+c_2z^2+c_4z^4+\cdots,\qquad(5\text{-}120)$$

对 $\nu>k$，我们规定 $a_\nu=0$，便得到

$$
\begin{aligned}
&b_0 = c_0 a_1,\\
&b_1 = c_0 a_2,\\
&b_{2\nu} = c_0 a_{2\nu+1} + c_2 a_{2\nu-1} + \cdots + c_{2\nu} a_1, \qquad (5\text{-}121)\\
&b_{2\nu+1} = c_0 a_{2\nu+2} + c_2 a_{2\nu} + \cdots\\
&\qquad\qquad + c_{2\nu} a_2,\ \ \nu = 1,2,\cdots.
\end{aligned}
$$

进一步的证明依赖由于(5-120)确定的系数 $c_{2\nu}$ 满足

$$
c_{2\nu} < 0,\ \ \nu = 1,2,\cdots. \qquad (5\text{-}122)
$$

为了不中断这个证明的主要思想，我们暂时不证明这个不等式而分二种情形来讨论.

(i) 若 k 为奇数,则由(5-121),有

$$
b_{k+1} = c_2 a_k + c_4 a_{k-2} + \cdots + c_{k+1} a_1.
$$

因为 $a_1 > 0$，a_ν 非负，所以由(5-122)即得 $b_{k+1} < 0$. 于是我们证明了:

定理 5.8. 一个步数 k 为奇数的稳定算子的阶不超过 $k+1$.

(ii) 若 k 为偶数,则

$$
b_{k+1} = c_2 a_k + c_4 a_{k-2} + \cdots + c_k a_2.
$$

鉴于我们已知道的 $c_{2\nu}$ 和 a_ν 的符号，$b_{k+1} = 0$ 的充分和必要条件则是 $a_2 = a_4 = \cdots = a_k = 0$. 这种情形当且仅当多项式 $r(z)$ 满足恒等式 $r(-z) = -r(z)$，即当 $r(z)$ 为奇数时才能成立. 由于 $r(z)$ 不能有实部为正的根,所以这个恒等式限制 $r(z)$ 在具有负实部的根上. 因此 $r(z)$ 的所有根必须位于虚轴上，这就是说，$\rho(\zeta)$ 的所有根位于 $|\zeta| = 1$ 上. 因为 $a_k = 0$，所以 $r(z)$ 的次数是 $k-1$，而 -1 是 $\rho(\zeta)$ 的一个根. 考虑到(5-115),(5-122)和 $a_1 > 0$，以及

$$
b_{k+2} = c_4 a_{k-1} + c_6 a_{k-3} + \cdots + c_{k+2} a_1
$$

为负值,便推得阶不超过 $k+2$. 于是我们就证明了:

定理 5.9. 一个稳定算子的阶 p 不超过 $k+2$. 对于

$p = k + 2$ 的一个充分必要条件是 k 为偶数，和 $\rho(\zeta)$ 的所有根模为 1，而 $\sigma(\zeta)$ 是由 (5-113) 确定的。

满足定理 5.9 条件的算子就称为最佳算子。定理 5.9 的证明提供了一些关于最佳算子的误差常数的有意义的不等式。由(5-104)，并根据 $\zeta - 1 = 2z + O(z^2)$，我们得到

$$p(z) = 2^{p-k} c_{p+1} z^p + O(z^{p+1}).$$

若 $p > k$ 则从 (5-119) 得到 $c_{p+1} = 2^{k-p} b_p$. 利用

$$\sigma(1) = 2^k s(0) = 2^k b_0,$$

我们得到

$$C = b_p / 2^p b_0. \tag{5-123}$$

因为一个最佳算子具有 $p = k + 2$，因此，利用 (5-121) 式，得到

$$C = \frac{c_4 a_{k-1} + c_6 a_{k-3} + \cdots + c_{k+2} a_1}{2^{k+1} c_0 a_1}.$$

再利用我们已知的 $c_{2\nu}$ 和 a_ν 的符号以及

$$c_0 = \frac{1}{2}, \quad c_4 = \frac{-2}{45},$$

便得到不等式

$$C \leqslant -\frac{1}{2^k \cdot 45} \frac{a_{k-1}}{a_1}, \tag{5-124}$$

$$C \leqslant 2^{-k-1} c_{k+2}. \tag{5-125}$$

因为 $c_{k+2} < 0$，所以 (5-125) 表明对给定的 k，$|C|$ 不能是任意小。

系数 $c_{2\nu}$ 的数值可以通过递推关系 $c_0 = \frac{1}{2}$，

$$c_{2\nu} + \frac{1}{3} c_{2\nu-2} + \frac{1}{5} c_{2\nu-4} + \cdots + \frac{1}{2\nu + 1} c_0 = 0,$$

$$\nu = 1, 2, \cdots$$

计算出来。表 5.12 中给出了其中某些数值。

表 5.12 (5-120) 中的系数 $C_{2\nu}$

ν	0	1	2	3	4
$C_{2\nu}$	$\frac{1}{2}$	$-\frac{1}{6}$	$-\frac{2}{45}$	$-\frac{22}{945}$	$-\frac{214}{14175}$

我们还必须证明不等式(5-122). 最快的方法便是把它作为由 Kaluza (1928) 得到如下结果的推论,对此我们只述而不证[1].

引理 5.4. 令 $f(t)=\sum_{\nu=0}^{\infty}A_\nu t^\nu$ 和 $g(t)=\sum_{\nu=0}^{\infty}B_\nu t^\nu$,

$$A_\nu>0(\nu=0,1,2,\cdots),\quad A_{\nu+1}A_{\nu-1}-A_\nu^2>0,$$

为满足 $f(t)g(t)=1$ 的二个幂级数,则

$$B_\nu<0\ (\nu=1,2,\cdots).$$

对于 $t=z^2$ 的函数,

$$f(t)=\frac{1}{z}\log\frac{1+z}{1-z}=2+\frac{2}{3}z^2+\frac{2}{5}z^4+\cdots.$$

容易证明它满足引理中的假设条件,因此对系数 $B_\nu=c_{2\nu}$ 结论是成立的.

5.2-9. 最佳算子的构造. 当 k 大时,由给定的多项式 $\rho(\zeta)$ 用 §5.2-7 中的方法构造多项式 $\sigma(\zeta)$ 则是不方便的. 下面介绍构造最佳算子的另一种方法.

假设多项式 $\rho(\zeta)$ 满足定理 5.9 中的条件, 又设与 ±1 相异的根分布于点 $e^{\pm i\varphi}$, $e^{\pm i\psi}$, $\cdots$, 则 $\rho(\zeta)$ 可以表示成

$$\rho(\zeta)=\alpha_k\zeta^{\frac{1}{2}k}(\zeta-\zeta^{-1})(\zeta-2\cos\varphi+\zeta^{-1})\times(\zeta-2\cos\psi+\zeta^{-1})\cdots.$$

从而

$$\rho(\zeta)=\zeta^{\frac{1}{2}k}(\zeta-\zeta^{-1})P(\zeta+\zeta^{-1}-2),\qquad(5\text{-}126)$$

---

1) 一个简单的证明可见 Amer. Math. Monthly, 66, 430, 1959.

其中 P 是 $\frac{1}{2}k-1$ 次多项式．若 $r(z)$ 为奇函数，则由(5-113)所确定的函数 $s(z)$ 是两个奇函数的乘积在 $z=0$ 处的 Taylor 多项式，故为偶函数．由 $s(-z)=s(z)$ 得到

$$\sigma(\zeta)=\zeta^k\sigma(\zeta^{-1})$$

或

$$\beta_{k-\nu}=\beta_\nu,\quad \nu=0,1,\cdots,k. \tag{5-127}$$

现在用归纳法很容易证明 $\sigma(\zeta)$ 可以表示成形式

$$\sigma(\zeta)=\zeta^{\frac{1}{2}k}Q(\zeta+\zeta^{-1}-2), \tag{5-128}$$

其中 Q 为一个 $\frac{1}{2}k$ 次多项式．[我们在(5-126)和(5-128)中选择变量为 $\zeta+\zeta^{-1}-2$ 而不用 $\zeta+\zeta^{-1}$，因为(5-58)式的右端项 $\nabla^{2\mu}f_{n+\frac{1}{2}k+\mu}$ 对应于 Q 中的项 $(\zeta+\zeta^{-1}-2)^\mu$]．现在我们令

$$\zeta+\zeta^{-1}-2=t^2, \tag{5-129}$$

则对于 $t=0$ 的邻域内的任何 t 与 $\zeta=1$ 的邻域内的两个值相对应．这两个值互为倒数，并且我们可以通过取公式

$$\zeta=\left(\sqrt{1+\frac{1}{4}t^2}+\frac{1}{2}t^2\right)^2$$

的平方根的两个可能的分支来得到它们的数值．我们通过选择，当 $t=0$ 时，平方根化成 $+1$ 的分支来定义单值函数 $\zeta(t)$．这样我们就有 $\zeta-\zeta^{-1}=2t\sqrt{1+\frac{1}{4}t^2}$，$\zeta-1=t+O(t^2)$．于是关系式(5-104)就等价于

$$\frac{t\sqrt{1+\frac{1}{4}t^2}}{\log\left(\frac{1}{2}t+\sqrt{1+\frac{1}{4}t^2}\right)}P(t^2)-Q(t^2)$$
$$=C_{k+3}t^{k+2}+O(t^{k+3}). \tag{5-130}$$

从而 $Q(t^2)$ 是与偶函数

$$\frac{t\sqrt{1+\frac{1}{4}t^2}}{\log\left(\frac{1}{2}t+\sqrt{1+\frac{1}{4}t^2}\right)}P(t^2)$$

$$= q_0 + q_2t^2 + q_4t^4 + \cdots \qquad (5\text{-}131)$$

的 k 阶 Taylor 多项式相等．因为 $\sigma(1) = Q(0) - q_0$，所以我们得到误差常数关系式

$$C = q_{k+2}/q_0. \qquad (5\text{-}132)$$

系数 $q_{2\nu}$ 可以通过多项式

$$P(t^2) = P_0 + P_2t^2 + \cdots + P_{k-2}t^{k-2} \qquad (5\text{-}133)$$

和函数

$$\frac{t\sqrt{1+\frac{1}{4}t^2}}{\log\left(\frac{1}{2}t+\sqrt{1+\frac{1}{4}t^2}\right)} = k_0 + k_2t^2 + k_4t^4 + \cdots$$

$$(5\text{-}134)$$

的 Taylor 级数的乘积求出． 按问题 23 中建立的递推关系，容易得到表 5.13 的数值．

表 5.13 (5-134) 中系数 $k_{2\nu}$

ν	0	1	2	3	4
$k_{2\nu}$	2	$\frac{1}{3}$	$-\frac{1}{90}$	$\frac{1}{753}$	$\frac{-6013}{7257600}$

例． (i) $k=2$ 的最佳方法．由于 $\rho(\zeta) = \zeta^2 - 1$，我们有 $P(t^2) = 1$，因此 $Q(t^2) = 2 + \frac{1}{3}t^2$． 项 $t^{2\nu}$ 是与 (5-58) 式右边的差分 $\nabla^{2\nu}f_{n+\frac{1}{2}k+\nu}$ 相对应的． 于是我们又导出 Milne 方法

$$y_{n+2}-y_n=h\left(2f_{n+1}+\frac{1}{3}\nabla^2 f_{n+2}\right).$$

(ii) $k=4$ 的最一般的最佳方法．设 $\rho(\zeta)$ 的根为

$$\zeta=\pm 1,\ \zeta=e^{\pm i\varphi},$$

则有

$$\rho(\zeta)=\zeta^2\left(\zeta-\frac{1}{\zeta}\right)\left(\zeta-2\cos\varphi+\frac{1}{\zeta}\right).$$

因此 $P(t^2)=4\lambda+t^2$，其中 $\lambda=\sin^2\frac{1}{2}\varphi$，$0<\lambda<1$．由

$$\frac{t\sqrt{1+\frac{1}{4}t^2}}{\log\left(\frac{1}{2}t+\sqrt{1+\frac{1}{4}t^2}\right)}P(t^2)=8\lambda+\left(2+\frac{4\lambda}{3}\right)t^2$$

$$+\left(\frac{1}{3}-\frac{2\lambda}{45}\right)t^4+\left(-\frac{1}{90}+\frac{\lambda}{189}\right)t^6+\cdots,$$

我们得到

$$Q(t^2)=8\lambda+\left(2+\frac{4\lambda}{3}\right)t^2+\left(\frac{1}{3}-\frac{2\lambda}{45}\right)t^4,$$

$$C=\frac{1}{1512}-\frac{1}{720\lambda}.$$

对于 $\lambda\to 1$，C 逼近于由 (5-125) 给出的上界 $2^{-5}c_6$．

iii) $k=6$ 的一个特殊最佳方法．多项式

$$\rho(\zeta)=\zeta^6-\zeta^5+\zeta^4-\zeta^2+\zeta-1$$

的根为 $\zeta=\pm 1$，$\zeta=e^{\pm i\pi/2}$，$\zeta=e^{\pm i\pi/3}$．因此

$$P(t^2)=\left(\zeta+\frac{1}{\zeta}\right)\left(\zeta+\frac{1}{\zeta}-1\right)=(t^2+2)(t^2+1).$$

于是我们得到多项式

$$Q(t^2)=4+\frac{20}{3}t^2+\frac{134}{45}t^4+\frac{286}{675}t^6,$$

并且所得到的方法可写成

$$y_{n+6}-y_{n+5}+y_{n+4}-y_{n+2}+y_{n+1}-y_n$$
$$=h\left(4f_{n+3}+\frac{20}{3}\nabla^2 f_{n+4}+\frac{134}{45}\nabla^4 f_{n+5}+\frac{286}{675}\nabla^6 f_{n+6}\right).$$

5.3. 线性多步方法的离散误差

5.3-1. 特殊情形下的离散误差. 下面所给出的离散误差处理尽可能与单步方法建立的模型一致. 不过，考虑到问题的复杂性，我们可以通过详细研究特殊问题

$$y'=Ay,\quad y(0)=1 \tag{5-135}$$

来着手讨论. 其中 A 是一个(实或复)常数，$A\neq 0$. 这里所得到的特殊结果在理论上也是需要的.

对于这个特殊问题 (5-135)，相应的差分方程 (5-58) 为

$$\alpha_k y_{n+k}+\alpha_{k-1}y_{n+k-1}+\cdots+\alpha_0 y_n$$
$$=Ah(\beta_k y_{n+k}+\beta_{k-1}y_{n+k-1}+\cdots+\beta_0 y_n). \tag{5-136}$$

这是常系数线性差分方程，它可以用 §5.2-1 中的方法求解. 其特征方程为 $\tilde{\rho}(\zeta)=0$,

$$\tilde{\rho}(\zeta)=\rho(\zeta)-Ah\sigma(\zeta), \tag{5-137}$$

这里 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 为 (5-81) 式所确定的多项式. 若 $\tilde{\rho}(\zeta)$ 的根 $\xi_1,\xi_2,\cdots,\xi_k$ 都是互异的，则根据定理 5.3，(5-136) 的解可以写成

$$y_n=\sum_{\mu=1}^{k}A_\mu\xi_\mu^n, \tag{5-138}$$

其中 A_μ 为适当的常数. 在用初始条件确定这些常数之前，我们来建立关于根 $\xi_\mu(\mu=1,\cdots,k)$ 的若干性质.

我们设由 (5-58) 定义的方法是稳定且相容的，此外还假

定多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 没有公共因子[1]. 显然，对于充分小的 h 值，根 ξ_μ 逼近于多项式 $\rho(\zeta)$ 的根 $\zeta_1, \zeta_2, \cdots, \zeta_k$ 中的某一个. 更确切地说，由复变理论的定理[2]推得，对于充分小的 $\varepsilon > 0$，存在 $\delta > 0$，使得对 $0 \leqslant h < \delta$，方程 $\bar{\rho}(\zeta) = 0$ 在每个圆 $|\zeta - \zeta_\mu| < \varepsilon(\mu = 1, 2, \cdots, k)$ 内根的个数恰好和

$$\rho(\zeta) = 0$$

的根在其中的个数相同. 如果 ζ_μ 是一个 p 重根，那么可以证明，p 个逼近于它的根 ξ_ν 是 $h^{1/p}$ 的解析函数的 p 个不同的值，而这 p 个值由 $h^{1/p}$ 给定. 由 $\sigma(\zeta_\mu) \neq 0$ 知，对于充分小的 h，$h \neq 0$，这 p 个根是不同的.

为了简化下面的讨论，我们给这些根 $\zeta_1, \cdots, \zeta_k$ 一个特殊的名称. 按稳定性假设，多项式 $\rho(\zeta)$ 没有模大于 1 的根. 如果有 m 个模为 1 的根，——再由稳定性假设，必为单根——用 $\zeta_1, \zeta_2, \cdots, \zeta_m$ 来表示. 特别是，我们可令 $\zeta_1 = 1$，根据相容性条件它总是一个根. 模为 1 的根称为本性根. 特别称 ζ_1 为主根. 如果有剩余的根 $\zeta_{m+1}, \cdots \zeta_k$，则称为非本性根.

我们需要对根 ξ_ν 作一些量的说明. 对逼近于非本性根 ζ_ν 的根，当 h 充分小时，下述粗糙的估计式

$$|\xi_\mu| \leqslant t, \quad \mu = m + 1, \cdots, k \tag{5-139}$$

是成立的，其中

$$t = \frac{1}{2}(1 + \max_{m < \mu \leqslant k} |\zeta_\mu|), \quad |t| < 1. \tag{5-140}$$

它表明对充分小的 h，近似于非本性根的根都位于包含在单位圆内的一个圆中.

---

1) §5.1 所讨论的特殊方法，都是满足这个假设的，从而由 §5.2-7 的讨论可见，$\rho(\zeta)$ 和 $\sigma(\zeta)$ 的公共因子可以消掉. 这既不改变方法的阶，又不改变方法的误差常数.

2) 例如，见 Ahlfors [1953], p.107, 定理 11.

根据上述，逼近于本性根的根为 h 的解析函数. 我们想通过它们以 h 为幂次的展开式来确定其线性项. 由待定系数法，我们得到

$$\xi_\mu = \zeta_\mu(1 + \lambda_\mu A h + O(h^2)), \quad \mu = 1, \cdots, m, \tag{5-141}$$

其中

$$\lambda_\mu = \frac{\sigma(\zeta_\mu)}{\zeta_\mu \rho'(\zeta_\mu)}. \tag{5-142}$$

常数 λ_μ 只依赖于方法而不依赖于所要积分的微分方程. 我们把这些 λ_μ 称之为增长参数，其原因在后面可以看到. 由相容性，我们注意到

$$\lambda_1 = \frac{\sigma(1)}{\rho'(1)} = 1. \tag{5-143}$$

我们感兴趣的是在 $nh = x$ 保持常数的意义下，当 $h \to 0$ 和 $n \to \infty$ 时 ξ_μ^n 的渐近性态. 考虑到

$$\begin{aligned}(1 + \lambda_\mu A h + O(h^2))^n &= \exp[x h^{-1} \log(1 + \lambda_\mu A h + O(h^2)]\\ &= \exp[x h^{-1}(\lambda_\mu A h + O(h^2))]\\ &= \exp(\lambda_\mu A x) + O(h)\end{aligned}$$

(也可见第一章中问题13)，令 $\zeta_\mu = e^{i\varphi_\mu}$，则

$$\xi_\mu^n = e^{in\varphi_\mu}[e^{\lambda_\mu A x} + O(h)], \quad h \to 0. \tag{5-144}$$

特别是，因为 $\zeta_1 = 1$，$\lambda_1 = 1$，所以

$$\xi_1^n = e^{Ax} + O(h). \tag{5-145}$$

我们希望更精确地确定出上面方程中的项 $O(h)$. 我们从计算 ξ_1 的一个更好的近似值着手考虑. 由关系式 (5-141) 得

$$\xi_1 = e^{Ah} + \tau,$$

其中 τ 为 $h = 0$ 处 h 的解析函数，$\tau = O(h^2)$. 代入(5-137)，就得到

$$\rho(e^{Ah}) + \rho'(e^{Ah})\tau + O(\tau^2) - hA\sigma(e^{Ah}) + O(h\tau) = 0.$$

我们利用在(5-58)中由 $y(x)=e^{Ax}$ 代入后得到的关系式

$$\rho(e^{Ah})-hA\sigma(e^{Ah})=C_{p+1}(hA)^{p+1}+O(h^{p+2}). \tag{5-146}$$

于是就有

$$\rho'(e^{Ah})\tau=-C_{p+1}(Ah)^{p+1}+O(h\tau)+O(h^{p+2}),$$

再利用 $\rho'(e^{Ah})=\rho'(1)+O(h)$,

$$\tau=-C(hA)^{p+1}+O(h^{p+2}),$$

其中 C 是由(5-93)确定的误差常数. 从而

$$\xi_1=e^{Ah}[1-(Ah)^{p+1}C+O(h^{p+2})]. \tag{5-147}$$

对于 $h\to0$, $nh=x$, 便得

$$\xi_1^n=e^{Ax}[1-xh^pA^{p+1}C+O(h^{p+1})]. \tag{5-148}$$

这就是我们所要的(5-145)更为精确的表达式.

现在我们假设已经给定了值 $y_0,y_1,\cdots,y_{k-1}$, 来确定(5-138)中的 A_μ. 为方便起见,我们把这些初始条件写成

$$y_\mu=\xi_1^\mu+\delta_\mu,\quad \mu=0,1,\cdots,k-1, \tag{5-149}$$

则 A_μ 所满足的方程为

$$A_1\xi_1^\mu+A_2\xi_2^\mu+\cdots+A_k\xi_k^\mu=\xi_1^\mu+\delta_\mu,$$
$$\mu=0,1,\cdots,k-1. \tag{5-150}$$

这组关于 k 个未知量 $A_1,\cdots,A_k$ 的线性方程可以按下面的方法求解. 对于 $\nu=1,2,\cdots,k$, 定义多项式

$$\tilde{\rho}_\nu(\zeta)=\frac{\tilde{\rho}(\zeta)}{\zeta-\xi_\nu}=\tilde{\alpha}_{\nu,0}+\tilde{\alpha}_{\nu,1}\zeta+\cdots+\tilde{\alpha}_{\nu,k-1}\zeta^{k-1}. \tag{5-151}$$

因为 $\tilde{\rho}(\xi_\mu)=0$, $\mu=1,\cdots,k$, 所以有

$$\tilde{\rho}_\nu(\xi_\mu)=\begin{cases}0, & \nu\neq\mu,\\ \tilde{\rho}'(\xi_\mu), & \nu=\mu.\end{cases} \tag{5-152}$$

用 $\tilde{\alpha}_{\nu,\mu}$ 乘第 μ 个方程(5-150), 相加, 并利用(5-152), 我们得到

$$A_\nu \tilde{\rho}'(\tilde{\xi}_\nu) = \tilde{\rho}_\nu(\tilde{\xi}_1) + \tilde{\Delta}_\nu, \tag{5-153}$$

其中

$$\tilde{\Delta}_\nu = \tilde{\alpha}_{\nu,0}\delta_0 + \tilde{\alpha}_{\nu,1}\delta_1 + \cdots + \tilde{\alpha}_{\nu,k-1}\delta_{k-1}.$$

因为对 $h \neq 0$ 时，考虑到 $\tilde{\xi}_1$ 是单根，所以 $\tilde{\rho}'(\zeta_\nu) \neq 0$. 从而

$$\begin{aligned} A_1 &= 1 + \frac{\tilde{\Delta}_1}{\tilde{\rho}'(\tilde{\xi}_1)}, \\ A_\nu &= \frac{\tilde{\Delta}_\nu}{\tilde{\rho}'(\tilde{\xi}_\nu)}, \quad \nu = 2, \cdots, k. \end{aligned} \tag{5-154}$$

当 $h \to 0$ 时，我们有 $\tilde{\Delta}_\nu = \Delta_\nu + O(h\delta)$，其中

$$\delta = \max|\delta_\mu|,$$

并且

$$\Delta_\nu = \alpha_{\nu,0}\delta_0 + \alpha_{\nu,1}\delta_1 + \cdots + \alpha_{\nu,k-1}\delta_{k-1}, \tag{5-155}$$

这里

$$\alpha_{\nu,0} + \alpha_{\nu,1}\zeta + \cdots + \alpha_{\nu,k-1}\zeta^{k-1} = \frac{\rho(\zeta)}{\zeta - \zeta_\nu}. \tag{5-156}$$

因为 $\tilde{\rho}'(\tilde{\xi}_\nu) = \rho'(\tilde{\xi}_\nu) - Ah\sigma'(\tilde{\xi}_\nu)$，若 ζ_ν 是一个 p 重根，则

$$\tilde{\rho}'(\tilde{\xi}_\nu) = \frac{1}{(p-1)!}\rho^{(p)}(\zeta_\nu)(\tilde{\xi}_\nu - \zeta_\nu)^{p-1} + O(h).$$

从而，若 $p = 1$，则 $\tilde{\rho}'(\tilde{\xi}_\nu) = \rho'(\zeta_\nu) + O(h)$；若 $p > 1$，则

$$\tilde{\rho}'(\tilde{\xi}_\nu) = O(h^{1-1/p}).$$

概括上述结果我们得到

$$\begin{aligned} y_n = &\left(1 + \frac{\Delta_1}{\rho'(1)} + O(h\delta)\right)e^{Ax}[1 - xA^{p+1}Ch^p + O(h^{p+1})] \\ &+ \sum_{\mu=2}^{m}\left(\frac{\Delta_\mu}{\rho'(\zeta_\mu)} + O(h\delta)\right)e^{in\varphi_\mu}[e^{\lambda_\mu Ax} + O(h)] \\ &+ \text{以 } h^{-1}t^n \text{ 为主的量}. \end{aligned}$$

由于 $h^{-1}t^n \to 0$ 比 h 的任何幂次都快，因此我们最终得到关于误差 $e_n = y_n - e^{Ax_n}$，只保留 $h \to 0$ 和 δ 的首阶项的渐近关

系式:

$$e_n = -CA^{p+1}xe^{Ax}h^p + O(h^{p+1}) + \sum_{\mu=1}^{m} \frac{\Delta\mu}{\rho'(\zeta_\mu)} e^{in\varphi_\mu} e^{\lambda_\mu Ax}$$
$$+ O(h\delta). \qquad (5-157)$$

这个公式表明离散误差实质上由两部分组成。第一部分由(5-157)中包含C的那项表示,称之为真正的离散误差.它恰好相当于具有主误差函数——$Cy^{(p+1)}(x)$的p阶单步方法中的离散误差。这个误差为x的光滑函数;此外还具有延迟趋向于该极限过程的特性。第二部分由(5-157)中的和表示,它的存在性是由于$\delta \neq 0$造成的,从而可以称之为开始误差。这个误差在性质上比真正的离散误差更复杂,需要几个注释。

(i) 如果$h \to 0$时$\delta \to 0$,则开始误差当$h \to 0$时显然趋于零。由于当$h \to 0$时真正的离散误差也趋于零,这就表明一个相容和稳定的多步方法对所考虑的特殊微分方程是收敛的(在§5.2-3定义的意义上)。

(ii) 如果开始值是精确的,即,若$y_\mu = e^{Ax_\mu}$, $\mu = 0, 1, \cdots, k-1$,则由(5-147),

$$\delta_\mu = e^{Ax_\mu} - \zeta_1^\mu = \mu C(Ah)^{p+1} + O(h^{p+2}).$$

于是,如果$Ah \neq 0$且$\mu > 0$,那么$\delta_\mu \neq 0$。从而开始误差不为零。不过它的阶数为$p+1$,即,比真正的离散误差小一个量级。

(iii) 如果$m = 1$,即,如果主根是$\rho(\zeta)$的唯一的本性根,则(5-157)中的和就化为它的第一项。此时开始误差与精确解成比例。从而在开始值稍有误差的情况下,用多步方法和用单步方法的结果是类似的。

(iv) 如果$m > 1$,且$\delta \neq 0$,则(5-157)的和中就出现了其它的项。这些项可以称为带有与$|e^{\lambda_\mu Ax}|$成比例的量的振动(由于因子$e^{in\varphi_\mu}$, $\varphi_\mu \neq 0$)。如果$\mathrm{Re}\,\lambda_\mu A \leqslant \mathrm{Re}\,A$,则这

些振动与精确解 e^{Ax} 比较是不明显的．但是，如果对某些 μ，$\operatorname{Re}\lambda_\mu A > \operatorname{Re} A$，那么这些振动对一个固定的 h 来说，与所要的精确解 e^{Ax} 比较将会随着 x 的增加而变大，从而使计算出来的值 y_n 变得无意义．例如，若 $A<0$ 并且增长参数中有一个是负的，就会出现这种情况．一切具有几个本性根的常用方法都有一些负的增长参数[1]． 作为一个例子，我们引用中点法则 $y_{n+1}-y_{n-1}=2hf_n$[(5-25)式，$q=0$ 的情形]． 这里我们有 $\zeta_2=-1$，$\lambda_2=-1$． 如果所要积分的微分方程是 $y'=-y$，且 $\delta\neq 0$，则由(5-157)得到，除了渐近于 $-\frac{1}{6}h^2xe^{-x}$ 的真正的离散误差外，还存在包含项 e^{-x} 和 $(-1)^ne^x$ 的线性组合的开始误差．第一项并不危险，然而第二项相对于精确解而言类似于 e^{2x} 那样增长． 从而对充分大的 x 值，y_n 的数值受到破坏．

表 5-14 中，我们给出了用步长 $h=0.1$ 计算的数值．虽然只给出了五位小数，但是这些数值都是以十位小数来计算的，且开始值 $y_1=e^{-h}$ 精确到十位小数．误差的振动性态一开始就是明显的，并且立刻呈现出使得数值无效的值．为了比较起见，还给出了用二阶 Adams-Bashforth 方法得到的近似值． 对于这个方法，因为 $m=1$，所以我们指望在误差中没有振动的成分．这由数值的光滑性得到证实．

对于 Milne 公式(5-33)，我们得到

$$\lambda_1=1,\ \lambda_2=-\frac{1}{3}.$$

因此在上面的例子中的开始误差包含 e^{-x} 和 $(-1)^ne^{x/3}$． 其

---

1) 在 §5.4-6 中将要证明对最佳差分算子(见 §5.2-8)对应于 $\zeta=-1$ 的增长参数总是小于 $-\frac{1}{3}$．

表 5.14

中点法则 Adams-Bashforth

x_n	y_n	e_n	y_n	e_n
0.0	1.00000	0.00000	1.00000	0.00000
0.1	0.90484	0.00000	0.90484	0.00000
0.2	0.81903	0.00030	0.81911	0.00038
0.3	0.74103	0.00021	0.74149	0.00070
0.4	0.67083	0.00051	0.67122	0.00090
⋮				
3.0	0.05152	0.00173	0.05043	0.00064
3.1	0.04363	−0.00142	0.04564	0.00060
3.2	0.04280	0.00204	0.04132	0.00056
3.3	0.03507	−0.00181	0.03740	0.00052
3.4	0.03578	0.00241	0.03386	0.00049
⋮				
5.0	0.01803	0.01129	0.00688	0.00015
5.1	−0.00775	−0.01385	0.00623	0.00014
5.2	0.01958	0.01406	0.00564	0.00012
5.3	−0.1166	−0.01665	0.00511	0.00011
5.4	0.02191	0.01739	0.00462	0.00011

次，“额外”分量的增长没有中点法则那么强，虽然与解本身比较还很强。

e_n 中的振动成分相对于精确解增长的现象通常称为数值不稳定． 这个不稳定和 §5.2-4 中讨论的不稳定不是同一类型的． 那个稳定性与 $\rho(\zeta)$ 的根模超过 1 有关． 具有这个性质的方法是发散的． 如果某个与本性根 ≈ 1 有关的增长参数是负的以及所要的解指数地减少，那么，在目前这一节中所讨论的现象就会在一个收敛方法中出现． 我们称这种现象为

弱稳定或条件稳定．$m=1$ 的收敛方法不会出现弱稳定，称之为强稳定．

上面的讨论虽然只是对特殊的单个微分方程进行的，但是这些结论也定性地适用于一般的初值问题．在舍入误差的讨论中，弱稳定和强稳定方法之间的区分也是主要的．

5.3-2. 两个引理．归根结底，对由线性多步法得到的任何微分方程的近似解的离散误差来建立类似于(5-157)的公式，是我们感兴趣的．这个工作并不象对单步法那样简单，并且需要大量的分析材料．在本节叙述的两个引理中的第二个引理在以后章节中将反复用到．第一个引理是因为第二个引理列公式和证明的需要．

引理 5.5. 设多项式 $\rho(\zeta)=\alpha_k\zeta^k+\alpha_{k-1}\zeta^{k-1}+\cdots+\alpha_0$ 满足稳定性条件，并令系数 $\gamma_l(l=0,1,2,\cdots)$ 由

$$\frac{1}{\alpha_k+\alpha_{k-1}\zeta+\cdots+\alpha_0\zeta^k}=\gamma_0+\gamma_1\zeta+\gamma_2\zeta^2+\cdots \tag{5-158}$$

确定，那么

$$\Gamma=\sup_{l=0,1,\cdots}|\gamma_l|<\infty.$$

证． 如果 $\hat\rho(\zeta)=\alpha_k+\alpha_{k-1}\zeta+\cdots+\alpha_0\zeta^k=\zeta^k\rho(\zeta^{-1})$，那么 $\hat\rho(\zeta)$ 的根是 $\rho(\zeta)$ 的根的倒数．因为 $\rho(\zeta)$ 没有在

$$|\zeta|=1$$

外部的根，所以 $1/\hat\rho(\zeta)$ 在 $|\zeta|<1$ 内是解析的，而且由于 $\rho(\zeta)$ 的根 $\zeta_1,\zeta_2,\cdots,\zeta_m$ 在 $|\zeta|=1$ 上都是单根，所以

$$[\hat\rho(\zeta)]^{-1} \text{ 在 } |\zeta|=1$$

上的极点是简单极点，因此存在常数 $A_1,A_2,\cdots,A_m$，使得函数 $f(\zeta)$

$$f(\zeta)=\frac{1}{\hat\rho(\zeta)}-\frac{A_1}{\zeta-\zeta_1^{-1}}-\cdots-\frac{A_m}{\zeta-\zeta_m^{-1}} \tag{5-159}$$

对于 $|\zeta| \leqslant 1$ 是解析的。根据 Cauchy 估计式（见 Alfors [1953], p.98）$f(\zeta)$ 在 $\zeta = 0$ 处的 Taylor 展开式的系数是有界的。由于每一项 $A_\mu/(\zeta - \zeta_\mu^{-1})$ 的展开式的系数也是有界的，从而引理得证。

我们需要恒等式

$$\alpha_k\gamma_l + \alpha_{k-1}\gamma_{l-1} + \cdots + \alpha_0\gamma_{l-k} = \begin{cases} 1, & l = 0, \\ 0, & l > 0, \end{cases} \tag{5-160}$$

其中假定 $l < 0$ 时，$\gamma_l = 0$。这个证明是通过用

$$\alpha_k + \alpha_{k-1}\zeta + \cdots + \alpha_0\zeta^k$$

乘以 (5-158) 式的两边，然后比较由此得到的以 ζ 为幂次的展开式的系数来完成的。

下面的引理与非齐次线性差分方程

$$\alpha_k z_{m+k} + \alpha_{k-1}z_{m+k-1} + \cdots + \alpha_0 z_m = h\{\beta_{k,m}z_{m+k} + \beta_{k-1,m}z_{m+k-1} + \cdots + \beta_{0,m}z_m\} + \lambda_m \tag{5-161}$$

解的增长有关。

引理 5.6. 令多项式 $\rho(\zeta) = \alpha_k\zeta^k + \cdots + \alpha_0$ 满足稳定性条件，令 B^*, β 以及 Λ 为非负常数，使得

$$|\beta_{k,n}| + |\beta_{k-1,n}| + \cdots + |\beta_{0,n}| \leqslant B^*,$$
$$|\beta_{k,n}| \leqslant \beta, |\lambda_n| \leqslant \Lambda, \quad n=0,1,2,\cdots,N. \tag{5-162}$$

又令 $0 \leqslant h < |\alpha_k|\beta^{-1}$，则 (5-161) 的每一个解当

$$|z_\mu| \leqslant Z, \quad \mu = 0,1,\cdots,k-1 \tag{5-163}$$

时满足

$$|z_n| \leqslant K^* e^{nhL^*}, \quad n = 0,1,\cdots,N, \tag{5-164}$$

其中

$$L^* = \Gamma^* B^*, \quad K^* = \Gamma^*(N\Lambda + AZK), \tag{5-165}$$

$$A = |\alpha_k| + |\alpha_{k-1}| + \cdots + |\alpha_0|, \Gamma^* = \frac{\Gamma}{1 - h|\alpha_k|^{-1}\beta}.$$

证. 对 $l=0,\cdots,n-k$，用 γ_l 乘相应于 $m=n-k-l$ 的方程（5-161）并把所得的方程相加. 设和为 S_n，则左边求和我们得到

$$S_n=(\alpha_k z_n+\alpha_{k-1}z_{n-1}+\cdots+\alpha_0 z_{n-k})\gamma_0$$
$$+(\alpha_k z_{n-1}+\alpha_{k-1}z_{n-2}+\cdots+\alpha_0 z_{n-k-1})\gamma_1+\cdots$$
$$+(\alpha_k z_k+\alpha_{k-1}z_{k-1}+\cdots+\alpha_0 z_0)\gamma_{n-k}.$$

重新整理后，得

$$S_n=\alpha_k\gamma_0 z_n+(\alpha_k\gamma_1+\alpha_{k-1}\gamma_0)z_{n-1}+\cdots$$
$$+(\alpha_k\gamma_{n-k}+\alpha_{k-1}\gamma_{n-k-1}+\cdots+\alpha_0\gamma_{n-2k})z_k$$
$$+(\alpha_{k-1}\gamma_{n-k}+\cdots+\alpha_0\gamma_{n-2k+1})z_{k-1}+\cdots$$
$$+\alpha_0\gamma_{n-k}z_0.$$

由（5-160），上式化成

$$S_n=z_n+(\alpha_{k-1}\gamma_{n-k}+\cdots+\alpha_0\gamma_{n-2k+1})z_{k-1}+\cdots$$
$$+\alpha_0\gamma_{n-k}z_0 \tag{5-166}$$

右边求和，我们得到

$$S_n=h\{\beta_{k,n-k}\gamma_0 z_n+(\beta_{k-1,n-k}\gamma_0+\beta_{k,n-k-1}\gamma_1)z_{n-1}+\cdots$$
$$+(\beta_{0,n-k}\gamma_0+\cdots+\beta_{k,n-2k}\gamma_k)z_{n-k}+\cdots+\beta_{0,0}\gamma_{n-k}z_0\}$$
$$+\lambda_{n-k}\gamma_0+\lambda_{n-k-1}\gamma_1+\cdots+\lambda_0\gamma_{n-k}. \tag{5-167}$$

使（5-116）和（5-167）相等，利用（5-162）和（5-163）求得

$$|z_n|\leqslant h\beta|\alpha_k^{-1}||z_n|+h\Gamma B^*\sum_{m=0}^{n-1}|z_m|+N\Gamma A+A\Gamma ZK.$$

对 $|z_n|$ 求解，就有

$$|z_n|\leqslant hL^*\sum_{m=0}^{n-1}|z_m|+K^*, \tag{5-168}$$

其中 L^* 和 K^* 由（5-165）给定. 现在我们用归纳法进行. 因为 $A\Gamma\geqslant 1$，从而 $K^*\geqslant Z$，估计式

$$|z_m|\leqslant K^*(1+hL^*)^m \tag{5-169}$$

对 $m=0,1,\cdots,k-1$ 是成立的. 设 $m=0,1,\cdots,n-1$

时上式成立，把它用于 (5-168) 的右边，我们得到

$$|z_n| \leqslant hL^*K^*\frac{(1+hL^*)^n-1}{hL^*}+K^*=K^*(1+hL^*)^n.$$

因此关系式 (5-169) 对 $m=n$ 是成立的，从而对于

$$m=0,1,\cdots,N$$

普遍成立。利用 $1+hL^*\leqslant e^{hL^*}$ 便得引理的结论。

在以后应用引理 5.6 时，常数 A, Z 和 N 通常依赖于 h。重要的是记住 Γ, A 以及在大多数情况下 B^* 和 β 都是不依赖于 h 的。

5.3-3. 收敛性的一个充分条件．本节我们将要证明，在 §5.2-4 和 §5.2-6 中得到的稳定性和相容性的条件不仅是线性多步方法收敛的必要条件，而且也是收敛的充分条件．

定理 5.10. 一个稳定且相容的线性多步方法是收敛的。

证． 假定函数 $f(x,y)$ 满足存在性定理 1.1 中的条件，又设 η 是一个任意常数．我们以 $y(x)$ 表示初值问题

$$y'=f(x,y),\quad y(a)=\eta$$

的解．而设 $y_n(n=0,1,2,\cdots)$ 为由开始值

$$y_\mu=\eta_\mu(h)\quad(\mu=0,1,\cdots,k-1)$$

所确定的差分方程 (5-58) 的解．令

$$\delta=\delta(h)=\max_{\mu=0,1,\cdots,k-1}|\eta_\mu(h)-y(a+\mu h)|,\tag{5-170}$$

并设

$$\lim_{h\to0}\delta(h)=0,\tag{5-171}$$

则我们必须证明对于任何 $x\in[a,b]$,

$$\lim_{\substack{h\to0\\x_n=x}}y_n=y(x).$$

我们先来估计量 $|L[y(x_m);h]|$，其中 L 表示由 (5-87) 所定义的差分算子．因为并未假设 $f(x,y)$ 是可微的，从而导数 $y''(x)$, $y'''(x)$, $\cdots$ 未必存在． 因此不能采用 §5.2-5 的方

法. 但是我们可以按下面的步骤来处理.

函数 $y'(x)=f(x,y(x))$ 在闭区间 $[a,b]$ 上是连续的. 对于 $\varepsilon\geqslant 0$, 我们定义量

$$\chi(\varepsilon)=\max_{\substack{|x^*-x|\leqslant\varepsilon\\ x,x^*\in[a,b]}}|y'(x^*)-y'(x)|,$$

对于 $\mu=0,1,2,\cdots,k$, 我们可以写成

$$y'(x_{m+\mu})=y'(x_m)+\theta_\mu\chi(\mu h),$$

其中 $|\theta_\mu|\leqslant 1$, 而且因为 $y(x_{m+\mu})=y(x_m)+\mu hy'(\zeta_\mu)$, 其中 $x_m<\zeta_\mu<x_{m+\mu}$, 我们有

$$y(x_{m+\mu})=y(x_m)+\mu h[y'(x_m)+\theta'_\mu\chi(\mu h)]$$

这里 $|\theta'_\mu|\leqslant 1$. 因而

$$\begin{aligned}L[y(x_m);h]&=(\alpha_0+\alpha_1+\cdots+\alpha_k)y(x_m)\\&+(\alpha_1+2\alpha_2+\cdots+k\alpha_k)y'(x_m)h\\&+\theta'(|\alpha_1|+2|\alpha_2|+\cdots+k|\alpha_k|)\chi(kh)h\\&-(\beta_0+\beta_1+\cdots+\beta_k)y'(x_m)h\\&-\theta(|\beta_0|+\cdots+|\beta_k|)\chi(kh)h\end{aligned}$$

其中 $|\theta|\leqslant 1$, $|\theta'|\leqslant 1$. 由于假设 L 是相容的, 所以,

$$\alpha_0+\alpha_1+\cdots+\alpha_k=0,$$

$\alpha_1+2\alpha_2+\cdots+k\alpha_k-\beta_0-\beta_1-\cdots-\beta_k=0$. 因此

$$|L[y(x_m);h]|\leqslant K\chi(kh)h, \tag{5-172}$$

其中

$$K=|\alpha_1|+2|\alpha_2|+\cdots+k|\alpha_k|+|\beta_0|+\cdots+|\beta_k|.$$

现在我们从值 y_m 所满足的相应的关系式

$$\alpha_ky_{m+k}+\cdots+\alpha_0y_m-h\{\beta_kf_{m+k}+\cdots+\beta_0f_m\}=0$$

中减去 $L[y(x_m);h]$. 记 $e_m=y_m-y(x_m)$, $m=0,1,\cdots$, 并令

$$g_m=\begin{cases}[f(x_m,y_m)-f(x_m,y(x_m))]e_m^{-1}, & e_m\neq 0,\\ 0, & e_m=0,\end{cases}$$

得到

$$\alpha_k e_{m+k} + \cdots + \alpha_0 e_m - h\{\beta_k g_{m+k} e_{m+k} + \cdots + \beta_0 g_m e_m\}$$
$$= \theta_m K \chi(kh) h,$$

其中 $|\theta_m| \leqslant 1$。 由于 Lipschitz 条件，$|g_m| \leqslant L$，$m = 0, 1, 2, \cdots$。 于是我们取 $z_m = e_m$，$z = \delta(h)$，$A = K\chi(kh)h$，$N = (x_n - a)/h$，$B^* = BL$，其中

$$B = |\beta_0| + |\beta_1| + \cdots + |\beta_k|,$$

便可应用引理 5.6。由此得到

$$|e_n| \leqslant \Gamma^*[A\delta(h) + (x_n - a)K\chi(kh)]$$
$$\times \exp[(x_n - a)L\Gamma^* B], \quad (5\text{-}173)$$

其中

$$A = |\alpha_0| + |\alpha_1| + \cdots + |\alpha_k|,$$
$$\Gamma^* = \frac{\Gamma}{1 - h|\alpha_k^{-1}\beta_k|L}. \quad (5\text{-}174)$$

因为 $y'(x)$ 在 $[a, b]$ 上一致连续，所以当 $h \to 0$ 时，

$$\chi(kh) \to 0.$$

据此由 (5-170) 可知，上面 e_n 的界对于任何 $x_n \in [a, b]$，随着 $h \to 0$ 而趋向零。这就证明了所要的结果。

我们注意到，如果收敛性按适当的方法定义，那么上述证明只要作一些变化，便能适用于 y 是复的以及 $f(x, y)$ 为复值的情形。在这种情况下，不必假定 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 的系数是实的；的确，引理 5.5 和 5.6 倘若没有这个假设也是成立的。

5.3-4. 关于离散误差的一个改进的先验界。界 (5-173) 一般来说是很粗糙的，而且，如果真正的解是充分光滑的，那么便不能显示出所期望的累积离散误差的真正量级。以 p 表示差分算子 (5-87) 的阶，又假定精确解 $y(x)$ 在 $[a, b]$ 上有 $p+1$ 阶连续导数，并令

$$Y = \max_{x \in [a,b]} |y^{(p+1)}(x)|. \quad (5\text{-}175)$$

我们将证明对于充分精确的开始值累积离散误差的阶为 h^p. 为此需要下面的引理.

引理 5.7. 设 $L[y(x);h]$ 为阶 $p>0$ 的差分算子，则存在一个只依赖于 L 的常数 $G>0$，使得对所有在 $[a,b]$ 上具有 $p+1$ 阶连续导数的函数 $y(x)$ 都有

$$|L[y(x);h]| \leqslant h^{p+1}GY, \quad a \leqslant x, \quad x+kh \leqslant b. \tag{5-176}$$

注. 由 §5.1-3 和 §5.1-4 的结果推出，对许多特殊的差分算子——即与 Adams 方法，Milne 方法，以及建立在微分法基础上有关的那些方法——我们可以写为

$$L[y(x);h] = h^{p+1}C_{p+1}y^{(p+1)}(\xi), \tag{5-177}$$

其中 C_{p+1} 由 §5.2-5 所确定，而 ξ 为区间 $(x, x+h)$ 内一个适当的数. 我们把 (5-177) 看作广义中值定理. 如果广义中值定理成立，那么令 $G=|C_{p+1}|$ 即得 (5-176). 但是对任意算子 (5-87)，并没有证明广义中值定理成立. 由 C_{p+1} 的定义，我们有

$$L[y(x);h] = h^{p+1}C_{p+1}y^{(p+1)}(x) + O(h^{p+2}),$$

而据此却不能推出 (5-176)，这是因为 (5-176) 并未考虑到项 $O(h^{p+2})$.

引理 5.7 的证明. 在一般情况下 G 的表达式可以利用余项为积分形式的 Taylor 定理得到. 如果 $y(x)$ 在 (a,b) 内有 $q+1$ 阶连续导数，且 x 与 $\bar{x}$ 在 (a,b) 内，那么我们有（见 Taylor [1955], p.112）

$$y(\bar{x}) = y(x) + (\bar{x}-x)y'(x) + \cdots + \frac{(\bar{x}-x)^q}{q!}y^{(q)}(x) + \frac{1}{q!}\int_x^{\bar{x}} (\bar{x}-t)^q y^{(q+1)}(t)\,dt.$$

令 $\bar{x} = x + \mu h (\mu = 1, 2, \cdots, k)$，把这个公式应用于 $y(x)$ 和

$y'(x)$，然后令 $t = x + hs$，我们得到

$$y(x+\mu h) = \cdots + h^{p+1}\int_0^{\mu}\frac{(\mu - s)^p}{p!}y^{(p+1)}(x+hs)ds,$$

$$y'(x+\mu h) = \cdots + h^{p+1}\int_0^{\mu}\frac{(\mu - s)^{p-1}}{(p-1)!}y^{(p+1)}(x+hs)ds,$$

其中这些点表示包含 $\leqslant p$ 阶导数的项。如果算子 $L[y(x);h]$ 为 p 阶，并且（5-87）右端的每一项都能用上面的表达式之一来代替，那么根据 p 的定义，这些由点表示的项都消失了，且余项可以写成

$$L[y(x);h] = h^{p+1}\int_0^{k} G(s)y^{(p+1)}(x+hs)ds. \tag{5-178}$$

函数 $G(s)$ 称为 $L[y(x);h]$ 的积分表达式的核。它在每个区间 $[\mu,\mu+1)(\mu = 0,1,\cdots,k-1)$ 中由不同的解析表达式来表示。对于 $s\in[k-1,k]$，我们有

$$G(s) = \alpha_k\frac{(k-s)^p}{p!} - \beta_k\frac{(k-s)^{p-1}}{(p-1)!},$$

而对于 $s\in[\mu,\mu+1)(\mu = 0,1,\cdots,k-2)$，

$$\begin{aligned}G(s) = {} & \alpha_k\frac{(k-s)^p}{p!} + \alpha_{k-1}\frac{(k-1-s)^p}{p!} + \cdots \\ & + \alpha_\mu\frac{(\mu+1-s)^p}{p!} - \beta_k\frac{(k-s)^{p-1}}{(p-1)!} - \cdots \\ & - \beta_\mu\frac{(\mu+1-s)^{p-1}}{(p-1)}.\end{aligned}$$

容易得到

$$|L[y(x);h]| \leqslant h^{p+1}\int_0^{k}|G(s)|\,|y^{(p+1)}(x+sh)|ds.$$

取

$$G = \int_0^{k}|G(s)|ds,$$

使得 (5-176) 式.

现在我们准备解决在本节开头所提出的问题. 为了后面的应用, 我们不假设序列 $\{y_n\}$ 是差分方程 (5-58) 的精确解, 而设 $\{y_n\}$ 是下列差分方程的解,

$$\alpha_k y_{m+k} + \alpha_{k-1} y_{m+k-1} + \cdots + \alpha_0 y_m - h\{\beta_k f_{m+k} + \cdots + \beta_0 f_m\} = \theta_m K_1 h^{q+1}, \quad m = 0,1,2,\cdots, \tag{5-179}$$

其中 K_1 和 q 都是非负常数, $|\theta_m| \leqslant 1$. 这样我们就可以证明:

定理 5.11. 在上面的假设下, 若 $|h\beta_k\alpha_k^{-1}|L < 1$, $x_n \in [a, b]$, 则

$$|e_n| \leqslant \Gamma^*[A\delta K + (x_n - a)(K_1 h^q + GYh^p)] \times \exp[(x_n - a)L\Gamma^* B], \tag{5-180}$$

其中 δ 是由 (5-170) 所确定的最大开始误差, 而 A 和 Γ 由 (5-174) 给出。

这个证明是定理 5.10 证明的简化形式. 由 (5-179) 减去量 $L[y(x_m);h]$, 我们得到

$$\alpha_k e_{m+k} + \cdots + \alpha_0 e_m - h\{\beta_k g_{m+k} e_{m+k} + \cdots + \beta_0 g_m e_m\} = \bar{\theta}_m[K_1 h^{q+1} + GYh^{p+1}], \quad m = 0,1,2,\cdots,$$

其中 $|\bar{\theta}_m| \leqslant 1$. 根据这个估计式, 再应用引理 5.6 即得 (5-180).

估计式 (5-180) 十分清楚地显示了局部误差的各种来源对累积误差的影响. 项 $A\delta K$ 表示开始误差的影响, 它实质上是不依赖于 h 的. 包含 K_1 和 GY 的项表示由局部不精确以及离散化所产生的误差. 这些误差都放大了 h^{-1} 阶的倍数.

建议读者对于某些特殊方法 (见问题 31) 确定 (5-180) 中出现的常数 A, B, G 和 Γ^* 的数值. 显然, 对于这样的一些方法, 如果常数 β_k 大且有交错符号(例如 Adams 方法), 那

么常数 B 就大，而这样对估计量是不利的．至于 Γ 的值，容易证明，对于所有建立在数值积分基础上的方法 $\Gamma = 1$． 因为这样的一些方法，$\rho(\zeta) = \zeta^k - \zeta^{k-q}$，其中 $1 \leqslant q \leqslant k$，所以 $\hat{\rho}(\zeta) = 1 - \zeta^q$，而

$$\frac{1}{\hat{\rho}(\zeta)} = \frac{1}{1 - \zeta^q} = 1 + \zeta^q + \zeta^{2q} + \cdots.$$

从而 $|\gamma_l| \leqslant 1$．对于显式方法我们有 $\beta_k = 0$，因此 $\Gamma^* = \Gamma$．

我们不加证明地指出，定理 5.11 的误差界也适用于 y 和 $f(x,y)$ 都是复的情形．在这种情况下，多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 未必为实系数．

5.3-5. 离散误差的渐近性态． 现在我们来研究对于固定的 x_n，当 $h \to 0$ 时，误差 e_n 的渐近性态．上一节的结果暗示出，在放宽条件的情况下，当 $h \to 0$ 时，$e_n \to 0$． 为了获得某些更有意义的结果，我们必须通过适当的 h 负幂次的放大来考察 e_n． 同样，为了得到适定的结果，必须符合一定的开始过程．我们假定开始值以及由此产生的开始误差都是 h 的充分可微的已知函数．我们可令

$$e_\mu = \delta_\mu(h), \quad \mu = 0, 1, \cdots, k-1,$$

并且规定 q 为使得

$$\delta_\mu = O(h^q), \quad \mu = 0, 1, \cdots, k-1 \qquad (5\text{-}181)$$

成立的最大整数． 假设 $q \geqslant 1$（例如，如果所有开始值是精确的，便有 $q = \infty$）． 此外，我们设 $y^{(p+2)}(x)$ 是连续的，并且函数 $y(x) = f_y(x, y(x))$ 在 $[a, b]$ 上是连续可微的． 最后，我们假设差分方程 (5-58) 是精确满足的，从而 (5-180) 中 $K_1 = 0$． 这样，由定理 5.11 推知 $e_n = O(h^r)$，这里

$$r = \min(p, q), \quad r \geqslant 1.$$

我们再一次从差分方程 (5-58) 减去相应的 $y(x)$ 满足的方程，写出余项为 $R = C_{p+1} y^{(p+1)}(x_n) h^{p+1} + O(h^{p+2})$． 利用

$$f(x_n, y_n) - f(x_n, y(x_n)) = g_n e_n + O(h^{2r}),$$
其中[1] $g_n = g(x_n)$，$g(x) = f_y(x, y(x))$． 我们得到

$$\alpha_k e_{n+k} + \alpha_{k-1} e_{n+k-1} + \cdots + \alpha_0 e_n = h\{\beta_k g_{n+k} e_{n+k} + \cdots + \beta_0 g_n e_n + O(h^{2r})\} - C_{p+1} y^{(p+1)}(x_n) h^{p+1} + O(h^{p+2}).$$

于是，对于伸缩误差 $\bar{e}_n = h^{-r} e_n$，我们得方程

$$\alpha_k \bar{e}_{n+k} + \alpha_{k-1} \bar{e}_{n+k-1} + \cdots + \alpha_0 \bar{e}_n = h\{\beta_k g_{n+k} \bar{e}_{n+k} + \cdots + \beta_0 g_n \bar{e}_n\} - C_{p+1} y^{(p+1)}(x_n) h^{p+1-r} + O(h^2) \qquad (5\text{-}182)$$

（我们用了 $2 \leqslant r+1$，$2 \leqslant p+2-r$）． 这个方程必须在初始条件

$$\bar{e}_\mu = h^{-r} \delta_\mu(h) \equiv \bar{\delta}_\mu(h), \quad \mu = 0, 1, \cdots, k-1 \qquad (5\text{-}183)$$

下来求解．

我们把解的形式表示成 $\bar{e}_n^I + \bar{e}_n^{II}$，其中 $\bar{e}_n^I$ 表示具有零初值的方程（5-182）的解，而 $\bar{e}_n^{II}$ 为对应的齐次方程

$$\alpha_k \bar{e}_{n+k} + \cdots + \alpha_0 \bar{e}_n = h\{\beta_k g_{n+k} \bar{e}_{n+k} + \cdots + \beta_0 g_n \bar{e}_n\} \qquad (5\text{-}184)$$

满足初始条件（5-183）的解．

我们先来确定 $\bar{e}_n^I$． 分两种情形讨论．若 $p > r$，则我们可应用引理 5.6，取 $z_m = \bar{e}_m^I$，$Z = 0$，$N = (b-a)/h$，$\beta_{\mu,m} = \beta_\mu g_{m+\mu}$，并用一个适当的常数 K，$\Lambda = Kh^2$． 于是便得

$$\bar{e}_n^I = O(h).$$

若 $p = r$，即开始误差不大于截断误差，则可以得到更有意义的结果．因为 $y^{(p+1)}(x)$ 在 $[a, b]$ 上有连续导数，所以我们可以写

$$C_{p+1} y^{(p+1)}(x_n) = C\{\beta_k y^{(p+1)}(x_{n+k}) + \cdots + \beta_0 y^{(p+1)}(x_n)\} + O(h).$$

1）g_n 的这个定义与 §5.3-4 中的稍有不同．

其中 $C = C_{p+1}/(\beta_k + \cdots + \beta_0)$ 是由 § 5.2-5 所确定的误差常数。于是 (5-182) 可以写成形式(略去 $\bar{e}^I$ 的上标)

$$\alpha_k\bar{e}_{n+k} + \cdots + \alpha_0\bar{e}_n = h\{\beta_k[y_{n+k}\bar{e}_{n+k} - Cy^{(p+1)}(x_{n+k})] + \cdots + \beta_0[y_n e_n - Cy^{(p+1)}(x_n)]\} + O(h^2).$$

假如用所讨论的多步方法解初值问题

$$e'(x) = g(x)e(x) - Cy^{(p+1)}(x), \quad e(a) = 0, \tag{5-185}$$

则得到同样的方程[没有 $O(h^2)$ 项]. 开始值 $\bar{e}^I_\mu = 0$ 与对应的值 $e(x_\mu)$ 至多相差 $O(h)$。 把定理 5.11 应用于现在的问题,得到

$$\bar{e}^I_n = e(x_n) + O(h), \tag{5-186}$$

此处 $e(x)$ 由 (5-185) 所确定.

现在我们来确定 $\bar{e}^H_n$。 我们把这个解表示为

$$\bar{e}^H_n = \sum_{\mu=1}^{k} A_\mu e_n^{(\mu)}, \tag{5-187}$$

其中序列 $\{e_n^{(\mu)}\}(\mu = 1,\cdots,k)$ 表示用下面的特殊方法定义的齐次方程 (5-184) 在 $n = 0$ 处的基本解组. 考察微分方程 $z'(x) = g_0 z(x)$。 如果用所考虑的多步方法解之,那么得到常系数线性差分方程

$$\alpha_k z_{n+k} + \cdots + \alpha_0 z_n = hg_0\{\beta_k z_{n+k} + \cdots + \beta_0 z_n\}. \tag{5-188}$$

如果 $g_0 \neq 0$, 且设 h 充分小但不为零,那么该差分方程具有形式为 ζ_μ^n 的基本解组 $z_n^{(\mu)}$, 而数 $\zeta_\mu(\mu = 1,\cdots,k)$ 是方程

$$\rho(\zeta) - hg_0\sigma(\zeta) = 0 \tag{5-189}$$

的根. 在这种情况下,我们用条件

$$e_\nu^{(\mu)} = \zeta_\mu^\nu, \quad \nu = 0,1,\cdots, k-1; \mu = 1,\cdots,k \tag{5-190}$$

来确定解 $e_n^{(\mu)}$。

如果 $g_0=0$，令 $s\leqslant k$ 为多项式 $\rho(\zeta)$ 的非零根的个数，则差分方程 (5-188) 有 s 个解．设 $z_n^{(\mu)}(\mu=1,2,\cdots,s)$ 是由定理 5.3 的构造所给出的基本组（若 $\rho(\zeta)$ 有多重非零根，它可含有如 $n\zeta^{n-1}$ 的解)，则我们用由

$$z_n^{(\mu)}=\begin{cases}1, & n=\mu-s-1,\\ 0, & \text{在其它情况下}\end{cases}$$

所确定的序列 $\{z_n^{(\mu)}\}(\mu=s+1,\cdots,k)$来补充这 s 个解．显然，当 $n\geqslant 0$ 时，它们也是 (5-188) 的解，并且矩阵

$$\begin{pmatrix} z_0^{(1)} z_1^{(1)} \cdots z_{k-1}^{(1)} \\ z_0^{(2)} z_1^{(2)} \cdots z_{k-1}^{(2)} \\ \cdots\cdots\cdots\cdots \\ z_0^{(k)} z_1^{(k)} \cdots z_{k-1}^{(k)} \end{pmatrix}$$

的行列式不为零．现在我们就用

$$e_\nu^{(\mu)}=z_\nu^{(\mu)},\ \nu=0,1,\cdots,k-1;\ \mu=1,\cdots,k \tag{5-191}$$

来确定解 $e_n^{(\mu)}$．

无论 $g_0=0$ 与否，都可以由 $n=0$ 处的基本组确定解 $e_n^{(\mu)}$，虽然它们以后可能会变为有关（因为我们并未假设 $n>0$ 时 $g_n\neq 0$)，但是由于定理 5.1，$e_n^{(\mu)}$ 能够用 (5-187) 的形式来表示．同样，如果我们用 $\xi_1,\cdots,\xi_m$ 表示 (5-189) 的那些根，它们逼近于 $\rho(\zeta)$ 的本性根 $\zeta_1=1,\zeta_2,\cdots,\zeta_m$，那么我们总有

$$e_\nu^{(\mu)}=z_\nu^{(\mu)}=\xi_\mu^\nu,\ \nu=0,1,\cdots,\ k-1;\ \mu=1,\cdots,m. \tag{5-192}$$

由 § 5.3–1 知存在常数 K 和 t，$|t|<1$，使得对所有充分小的 h 值，

$$|z_n^{(\mu)}|\leqslant Kt^n,\ \mu=m+1,\cdots,k;\ n=1,2,\cdots. \tag{5-193}$$

现在我们来建立解 $e_n^{(\mu)}$ 的渐近关系．将要证明对于 $\mu=$

$1,\cdots,m$ 这些解可以用单个微分方程的解来近似，而对于 $\mu>m$，它们类似于 (5-193) 那样迅速衰减。

对于 $\mu=1,\cdots,m$，令 $f_n^{(\mu)}=\zeta_\mu^{-n}e_n^{(\mu)}$。由 (5-184) 和 (5-192) 量 $f_n^{(\mu)}$ 满足

$$f_\nu^{(\mu)}=\zeta_\mu^{-\nu}\xi_\mu^\nu=1+O(h),\ \nu=0,1,\cdots,k-1,\tag{5-194}$$

$$\alpha_k^{(\mu)}f_{n+k}^{(\mu)}+\cdots+\alpha_0^{(\mu)}f_n^{(\mu)}=h\{\beta_k^{(\mu)}g_{n+k}f_{n+k}^{(\mu)}+\cdots+\beta_0^{(\mu)}g_nf_n^{(\mu)}\};\quad n=0,1,\cdots,\tag{5-195}$$

其中

$$\begin{aligned}\alpha_k^{(\mu)}\zeta^k+\cdots+\alpha_0^{(\mu)}&=\rho(\zeta_\mu\zeta)\equiv\rho^{[\mu]}(\zeta),\\ \beta_k^{(\mu)}\zeta^k+\cdots+\beta_0^{(\mu)}&=\sigma(\zeta_\mu\zeta)\equiv\sigma^{[\mu]}(\zeta).\end{aligned}\tag{5-196}$$

如果 $\sigma^{[\mu]}(\zeta)$ 被 §5.3-1 [方程 (5-142)] 确定的增长参数 λ_μ 去除，那么由多项式 $\rho^{[\mu]}(\zeta)$ 和 $\sigma^{[\mu]}(\zeta)$ 确定的线性多步方法是稳定且相容的。因为由 (5-196) 得

$$\rho^{[\mu]}(1)=\rho(\zeta_\mu)=0,$$
$$\rho^{[\mu]\prime}(1)=\zeta_\mu\rho'(\zeta_\mu)=\lambda_\mu^{-1}\sigma(\zeta_\mu)=\lambda_\mu^{-1}\sigma^{[\mu]}(1).$$

因此，如果我们用 λ_μ 除 (5-195) 中的 $\beta_k^{(\mu)}$ 以及用 λ_μ 乘量 g_m，则差分方程 (5-195) 变成对微分方程

$$e_\mu'(x)=\lambda_\mu g(x)e_\mu(x)\tag{5-197a}$$

的相容且稳定的逼近。而且初值 (5-194) 与满足

$$e_\mu(a)=1\tag{5-197b}$$

的 (5-197a) 之解的对应值 $e_\mu(x_\nu)$ 至多相差 $O(h)$。

现在我们要援引定理 5.11 [取 $p=1,\ k=0,\delta=O(h)$] 及其下面证明的注解得到

$$f_n^{(\mu)}=e_\mu(x_n)+O(h),\ \mu=1,\cdots,m;\ n=0,1,\cdots,$$

其中 $e_\mu(x)$ 由 (5.197) 确定。从而

$$e_n^{(\mu)}=e^{in\varphi_\mu}e_\mu(x_n)+O(h).\tag{5-198}$$

这个关系式解决了属于本性根的解 $e_n^{(\mu)}$ 的渐近性态。为

了证明其余的解当 $h\to 0$ 时趋于零，由引理 5.6 [采用 $z_m=e_m^{(\mu)}$, $Z=1$, $N=(b-a)/h$, $\beta_{\mu,m}=\beta_\mu g_{m+\mu}$, $\Lambda=0$]，我们首先注意到

$$e_n^{(\mu)}=O(1)\quad h\to 0. \tag{5-199}$$

为了改进这个结果，我们令 $e_n^{(\mu)}=z_n^{(\mu)}+\delta_n$，由 (5-191) 可见

$$\delta_\nu=0,\ \nu=0,\ 1,\cdots,k-1. \tag{5-200}$$

我们又令 $g_n=g_0+\varepsilon_n$. 代入 (5-184)，得到

$$\begin{aligned}\alpha_k(z_{n+k}^{(\mu)}+\delta_{n+k})+\cdots+\alpha_0(z_n^{(\mu)}+\delta_n)\\=hg_0\{\beta_k(z_{n+k}^{(\mu)}+\delta_{n+k})+\cdots+\beta_0(z_n^{(\mu)}+\delta_n)\}\\+h\{\beta_k\varepsilon_{n+k}e_{n+k}^{(\mu)}+\cdots+\beta_0\varepsilon_n e_n^{(\mu)}\}.\end{aligned}$$

由 (5-198) 式可知，含有 $z_m^{(\mu)}$ 的项消去了. 由于 (5-199) 以及 $g(x)$ 是可微的，最后一行可以简化. 因此我们得到

$$\begin{aligned}\alpha_k\delta_{n+k}+\cdots+\alpha_0\delta_n=hg_0\{\beta_k\delta_{n+k}+\cdots+\beta_0\delta_0\}\\+hC\theta_n\varepsilon_n.\end{aligned} \tag{5-201}$$

其中 C 是一个不依赖于 h 的适当常数和 $|\theta_n|\leqslant 1$. 现在我们应用引理 5.6，令 $Z_m=\delta_m$, $Z=0$ [根据 (5-200) 这是可能的]，$\beta_{\mu,m}=g_0\beta_\mu$,

$$N=\frac{2\log h}{|\log t|}=l \tag{5-202}$$

[这里 t 由 (5-140) 给出]，并且因为 $|\varepsilon_n|\leqslant nhG$，如果

$$G=\max|g'(x)|,\ \Lambda=CGlh^2$$

的话. 由此推知，对于 $n\leqslant l+k$, $|\delta_n|\leqslant C_1(h\log h)^2$，其中 C_1 是一个适当的常数. 另一方面，对于 $n\geqslant l$，由 (5-193) 和 (5-202) 得出 $|z_n^{(\mu)}|\leqslant Kt^n\leqslant Kh^2$. 因此对于 $l\leqslant n\leqslant l+k$，我们便有

$$|e_n^{(\mu)}|=|z_n^{(\mu)}+\delta_n|\leqslant Kh^2+C_1(h\log h)^2\leqslant K_1h^{2-\varepsilon}, \tag{5-203}$$

其中 ε 是一个任意正数. 于是证明了，当 $h\to 0$ 时，使得

$nh \to 0$ 的 k 个相邻的值 n，$e_n^{(\mu)}$ 为 $O(h^{2-\varepsilon})$. 再一次把引理 5.6 应用于 $n = l$ 开始的差分方程，于是最后的估计式（用一个不同的常数 K_1）对于 $n \geq l$ 只要 $x_n \in [a, b]$ 仍然成立。

剩下的问题是确定 (5-187) 中的常数 A_μ，把它作为开始值的函数。因为 $e_\nu^{(\mu)} = z_\nu^{(\mu)}$，$\nu = 0, 1, \cdots, k-1$，为了确定 A_μ，我们可以用 $z_n^{(\mu)}$ 来代替 $e_n^{(\mu)}$. 而对于 $z_n^{(\mu)}$，类似的计算在 §5.3-1 中已经完成。对于 $\mu \geq m$，只要确定 A_μ 的量级就够了。由指数 q 的定义，开始值本身是 $O(1)$. 如果 $g_0 \neq 0$，那么由 (5-154) 得出 $A_\mu = O(h^{-1+1/p_\mu})$，其中 p_μ 是当 $h \to 0$ 时 ξ_μ 逼近于根 ζ_μ 的重数。如果 $g_0 = 0$，则所有的 A_μ 为 $O(1)$. 因为我们可以在 (5-203) 中选取

$$\varepsilon > 1/p_\mu \quad (\mu = m+1, \cdots, k),$$

所以当 $x_n = x$，$a < x \leq b$ 时，两种情况都有

$$A_\mu e_n^{(\mu)} = O(h), \quad \mu = m+1, \cdots, k.$$

当 $\mu = 1, 2, \cdots, m$ 时，系数 A_μ 可以象 §5.3-1 那样完全确定. 令

$$\Delta_\mu = \lim_{h\to 0} h^{-r}\{\alpha_{\mu,0}\delta_0(h) + \alpha_{\mu,1}\delta_1(h) + \cdots + \alpha_{\mu,k-1}\delta_{k-1}(h)\}, \tag{5-204}$$

我们求得

$$A_\mu = \frac{\Delta_\mu}{\rho'(\zeta_\mu)} + O(h), \quad \mu = 1, \cdots, m.$$

因此，最终有

$$\tilde{e}_n^H = \sum_{\mu=1}^{m} \frac{\Delta_\mu}{\rho'(\zeta_\mu)} e^{in\varphi_\mu} e_\mu(x_n) + O(h).$$

现将本节结果概括如下：

定理 5.12. 在 §5.3-5 开始时所叙述的假设下，当 $h \to 0$，$nh = x - a$，$a < x \leq b$ 时由线性多步方法 (5-58) 得到的

$y' = f(x, y)$ 的解的离散误差渐近性态为

$$e_n = h^p e(x) + h^r \sum_{\mu=1}^{m} \frac{\Delta_\mu}{\rho'(\zeta_\mu)} e^{in\varphi_\mu} e_\mu(x) + O(h^{r+1}), \tag{5-205}$$

其中 $r = \min(p, q)$，而 Δ_μ 是由 (5-204) 所给出的开始值函数。 函数 $e(x)$ 和 $e_\mu(x)$ 分别是由初值问题 (5-185) 和 (5-197) 所确定的解。

这个结果是 (5-157) 的直接推广。 误差又由两部分组成。 真正的离散误差由包含 $e(x)$ 的项表示。 它恰好对应于主误差函数为 $-Cy^{(p+1)}(x)$的单步方法的离散误差。 开始误差由包含函数 $e_\mu(x)$ 的和表示。 如果 $m > 1$，那么，一般来说，连开始误差展开式的首项都不是 x 的光滑函数。因此，如果 $q \leqslant p$（即，若开始值相对精确的话），那么e_n的渐近展式的首项是不光滑的，并且外推到极限的方法不能够严格应用[1]。而且，在 $g(x) < 0$ 的指数递减解的近似中，如果对某个 μ，$\lambda_\mu < 0$，则对应的函数 $e_\mu(x)$ 将指数地增加，并且由于因子 $e^{in\varphi_\mu}$ 的迅速振动而使它本身产生明显的振动。这种振动即使 $q > p$（即，若开始误差比截断误差小）也会成为某些麻烦。另一方面，如果 $m = 1$（如所有 Adams 方法），则(5-181)的首项即使开始值不精确也总是 x 的光滑函数。这种自稳定的性质代表了强稳定方法的主要优点之一。

5.3-6. 估计局部离散误差的 Milne 方法. 如果离散误差主要包含真正的离散误差 [如在 (5-205)中，$p < q$ 的情形]，那么我们通过积分由函数$e(x)$确定的微分方程可以得到该误差大小的初步概念。这就必须知道，至少近似地知道，起着主误差函数作用的函数$-Cy^{(p+1)}(x)$。 即使不要积分 (5-185)，

1) 见 Wasow [1955] 中关于外推到极限不能起作用的有关例子。

该函数也是可以估价的，例如，可以从确定所用步长是否适当起见作估价. 用 $f(x, y)$ 的 p 阶导函数来确定 $y^{(p+1)}(x)$ 通常是不实用的. W. E. Milne 提出的方法可以用一种十分简单的方法得到 $y^{(p+1)}(x)$ 的近似值. 它是建立在 y_{n+k} 的预估值与最终接受的“校正”值相比较的基础上的[1].

我们首先在简化假设，即值 $y_n, y_{n-1}, \cdots, y_{n+k-1}$ 与精确解 $y(x)$ 的值一致的前提下来叙述 Milne 方法. 设预估公式为

$$\alpha_k^* y_{n+k}^* + \alpha_{k-1}^* y_{n+k-1} + \cdots + \alpha_0^* y_n = h\{\beta_{k-1}^* f_{n+k-1} + \cdots + \beta_0^* f_n\}, \qquad (5\text{-}206\text{p})$$

并设校正公式为

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = h\{\beta_k f_{n+k} + \beta_{k-1} f_{n+k-1} + \cdots + \beta_0 f_n\}. \qquad (5\text{-}206\text{c})$$

这里不假定 $|\alpha_0| + |\beta_0| > 0$. 但重要的是预估公式和校正公式的阶数相同. 因此，如果 $z(x)$ 是任意充分可微函数，那么我们有

$$\alpha_k^* z(x+kh) + \cdots + \alpha_0^* z(x) = h\{\beta_{k-1}^* z'(x+(k-1)h + \cdots + \beta_0^* z'(x)\} + C_{p+1}^* z^{(p+1)}(x) h^{p+1} + O(h^{p+2}), \qquad (5\text{-}207\text{p})$$

$$\alpha_k z(x+kh) + \cdots + \alpha_0 z(x) = h\{\beta_k z'(x+kh) + \cdots + \beta_0 z'(x)\} + C_{p+1} z^{(p+1)}(x) h^{p+1} + O(h^{p+2}), \qquad (5\text{-}207\text{c})$$

其中 C_{p+1}^* 和 C_{p+1} 是两个分别依赖于 $\alpha_k^*, \cdots, \beta_0^*$ 和 $\alpha_k, \cdots, \beta_0$ 的非零常数. 实际上，如果我们选取 $x = x_n$, $z(x) = y(x)$ 以及从 (5-206p) 和 (5-207c) 减去 (5-207p) 并考虑到

$$y_m = y(x_m) \quad (m = n, n+1, \cdots, n+k-1),$$

我们便得到

$$\alpha_k^*(y_{n+k}^* - y(x_{n+k})) = -C_{p+1}^* y^{(p+1)}(x_n) h^{p+1} + O(h^{p+2}), \qquad (5\text{-}208\text{p})$$

---

1) 关于预估-校正方法过程，请参阅 §5.1-2(ii).

$$\alpha_k(y_{n+k}-y(x_{n+k}))=h\beta_k[f(x_{n+k},y_{n+k})-f(x_{n+k},y(x_{n+k}))]$$
$$-C_{p+1}y^{(p+1)}(x_n)h^{p+1}+O(h^{p+2}),\qquad(5\text{-}208c)$$

方括弧中的表达式可以写成 $f_y(x_{n+k},\eta)(y_{n+k}-y(x_{n+k}))$. 求解 $y_{n+k}-y(x_{n+k})$ 便得

$$\alpha_k(y_{n+k}-y(x_{n+k}))=-C_{p+1}y^{(p+1)}(x_n)h^{p+1}$$
$$+O(h^{p+1}).\qquad(5\text{-}208c')$$

从 (5-208p) 和 (5-208c) 中消去 $y(x_{n+k})$ 并解出 $y^{(p+1)}(x_n)$, 我们得到

$$y^{(p+1)}(x_n)=h^{-p-1}\left(\frac{C^*_{p+1}}{\alpha^*_k}-\frac{C_{p+1}}{\alpha_k}\right)(y_{n+k}-y^*_{n+k})$$
$$+O(h).\qquad(5\text{-}209)$$

该公式是用预估值 y_{n+k} 和校正值 y^*_{n+k} 之差来表示未知导数 $y^{(p+1)}$ (近似地)的.

为了证明在一般情况下可以用同样的方法计算 $y^{(p+1)}$, 这里 $y_n,y_{n+1},\cdots,y_{n+k-1}$ 并未假定是精确的,必须假设预估值和校正公式的左端是相同的,即, $\alpha^*_\mu=\alpha_\mu,\mu=0,\cdots,k$. 而且,校正公式和开始值必须使得

$$e_n=y_n-y(x_n)=h^pe(x_n)+O(h^{p+1}),\quad(5\text{-}210)$$

其中 $e(x)$ 是 x 的连续可微函数[不需要由(5-185)确定 $e(x)$]且 $p\geqslant1$. 在(5-205)中,如果 $q>p$, 或者 $q=p$ 以及 $m=1$, 则该条件是满足的. 在这些假设下, 我们在 (5-207p) 中令 $x=x_n$, $z(x)=y(x)$ 并且从 (5-206p) 中减去由此所得的关系式,得到

$$\alpha_ke^*_{n+k}+\alpha_{k-1}e_{n+k-1}+\cdots+\alpha_0e_n$$
$$=h\{\beta^*_{k-1}[f(x_{n+k-1},y_{n+k-1})-f(x_{n+k-1},y(x_{n+k-1}))]+\cdots$$
$$+\beta^*_0[f(x_n,y_n)-f(x_n,y(x_n))]\}$$
$$-h^{p+1}C^*_{p+1}y^{(p+1)}(x_n)+O(h^{p+2}),\qquad(5\text{-}211)$$

其中 $e^*_{n+k}=y^*_{n+k}-y(x_{n+k})$. 根据 (5-210), 方括弧中的表

达式可以用

$$h^p g(x_m)(e(x_m)+O(h)),\quad m=n,n+1,\cdots,n+k-1$$

来代替，其中 $g(x)=f_y(x,y(x))$。而且，因为假定 $e(x)$ 是连续可微的，

$$g(x_{n+\mu})e(x_{n+\mu})=g(x_n)e(x_n)+O(h),$$
$$\mu=1,\cdots,k-1.$$

所以 (5-211) 最终可以写成形式

$$\alpha_k e^*_{n+k}+\alpha_{k-1}e_{n+k-1}+\cdots+\alpha_0 e_n$$
$$=h^{p+1}\{(\beta^*_k+\cdots+\beta^*_0)g(x_n)e(x_n)-C^*_{p+1}y^{(p+1)}(x_n)\}$$
$$+O(h^{p+2}). \tag{5-212p}$$

用完全类似的方法，由 (5-207c) 和 (5-208c)，我们得到

$$\alpha_k e_{n+k}+\alpha_{k-1}e_{n+k-1}+\cdots+\alpha_0 e_0$$
$$=h^{p+1}\{(\beta_k+\cdots+\beta_0)g(x_n)e(x_n)-C_{p+1}y^{(p+1)}(x_n)\}$$
$$+O(h^{p+2}). \tag{5-212c}$$

现在我们从 (5-212c) 减去 (5-212p) 并由相容性知，

$$\beta^*_{k-1}+\cdots+\beta^*_0=k\alpha_k+(k-1)\alpha_{k-1}+\cdots+\alpha_1$$
$$=\beta_k+\cdots+\beta_0.$$

因此，由 $e_{n+k}-e^*_{n+k}=y_{n+k}-y^*_{n+k}$，我们得到

$$\alpha_k(y_{n+k}-y^*_{n+k})=h^{p+1}(C^*_{p+1}-C_{p+1})y^{(p+1)}(x_n)$$
$$+O(h^{p+2}).$$

容易解出主误差函数的值为

$$-Cy^{(p+1)}(x_n)=h^{-p-1}K(y_{n+k}-y^*_{n+k})+O(h), \tag{5-213}$$

其中常数

$$K=\frac{C\alpha_k}{C_{p+1}-C^*_{p+1}}$$

仅依赖于所用的预估和校正公式．引进与预估公式有关的误差常数

$$C^* = \frac{C_{p+1}^*}{\beta_0 + \cdots + \beta_k},$$

我们也可以写成

$$K = \frac{\alpha_k}{\beta_0 + \cdots + \beta_k} \frac{C}{C - C^*}. \qquad (5\text{-}214)$$

于是我们证明了:

定理 5.13. 如果近似解满足 (5-210), 预估和校正公式是同阶的, 并且与它们有关的多项式 $\rho(\zeta)$ 是相同的(相乘一个适当的 ζ 幂次以后是可能的), 那么对任何固定的 $x \in [a, b]$, 当 $h \to 0$, $x_n = x$ 时, (5-213) 成立.

表 5.15 (5.24) 中常数 K 的数值

阶 p		1	2	3	4	5	6
K	Adams-Bashforth Moulton	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{10}$	$\frac{19}{270}$	$\frac{27}{502}$	$\frac{863}{19950}$
	Nyström-Milne Simpson	$-\frac{1}{2}$	∞	0	$\frac{1}{60}$	$\frac{1}{58}$	$\frac{37}{2352}$

例如, Adams-Moulton 方法, 或 Nystöm 和 Milne-Simpson 方法都是合适的一对预估-校正公式. 在表 5.15 中, 我们对不同的阶数 p 列出这两对公式的常数 K 的值.

后面一排数值所以不规则, 是由于 $p = 2$ 时 Nyström 和 Milne-Simpson 公式恰好相合, 以及 $p = 3$ 时 Milne-Simpson 公式自动变为 4 阶引起的. 应当注意, $p \geqslant 3$ 时 K 的数值非常小 (<0.1), 因此, 局部离散误差 $-Ch^{p+1}y^{(p+1)}(x)$ 仅仅是校正值和预估值之差的一个小的部分. 为了表示局部离散误差以及防止出错, Milne 建议在一个特殊的"检查"列中计算 $K(y_{n+k} - y_{n+k}^*)$ 的值.

5.3-7. 仅用一次校正的方法. 如果确定 y_{n+k} 的方程是隐式的(即, 如果 $\beta_k \neq 0$), 那么我们直到现在总是假定(5-58)

是精确满足的．如果这个方程可以用 §5.2-2 所叙述的迭代法求解，那么，一般来说需要迭代无穷多次．当然，实际上只能进行有限次的迭代，因此就不能精确满足 (5-58)． 由此产生的误差可以看成局部舍入误差，并由 §5.4 中将要讨论的方法来处理．考察同一个问题的另一种方法是对所要完成的迭代一次固定．这是一种新的方法，一般不再用线性多步方法求解微分方程．下面我们详细考虑只进行一次迭代的情形．

设预估公式为

$$\begin{aligned} y^*_{n+k} &+ \alpha^*_{k-1} y_{n+k-1} + \cdots + \alpha^*_0 y_n \\ &= h\{\beta^*_{k-1} f_{n+k-1} + \cdots + \beta^*_0 f_n\}, \qquad (5\text{-}215) \end{aligned}$$

并设它是 q 阶的，于是对于充分可微的函数 $z(x)$，就有

$$\begin{aligned} z(x+kh) &+ \alpha^*_{k-1} z(x+(k-1)h) + \cdots + \alpha^*_0 z(x) \\ &= h\{\beta^*_{k-1} z'(x+(k-1)h) + \cdots + \beta^*_0 z'(x)\} \\ &\quad + C^*_{q+1} z^{q+1}(x) h^{q+1} + O(h^{q+2}). \qquad (5\text{-}216) \end{aligned}$$

为了后面代数的简化，我们设 $\alpha^*_k = 1$． 校正公式取通常的形式

$$\alpha_k y_{n+k} + \cdots + \alpha_0 y_n = h\{\beta_k f_{n+k} + \cdots + \beta_0 f_n\},$$

设它的阶为 p，这里不需要 $p = q$，于是

$$\begin{aligned} \alpha_k z(x+kh) &+ \cdots + \alpha_0 z(x) \\ &= h\{\beta_k z'(x+kh) + \cdots + \beta_0 z'(x)\} \\ &\quad + C_{p+1} z^{(p+1)}(x) h^{p+1} + O(h^{p+2}). \qquad (5\text{-}217) \end{aligned}$$

这样，只用一次校正的算法可以叙述如下．假设

$$y_{n+k-1}, \cdots, y_n$$

的值是已知的，则由 (5-215) 计算 y^*_{n+k}，而由

$$\begin{aligned} \alpha_k y_{n+k} &+ \cdots + \alpha_0 y_n = h\{\beta_k f(x_{n+k}, y^*_{n+k}) \\ &+ \beta_{k-1} f(x_{n+k-1}, y_{n+k-1}) + \cdots + \beta_0 f(x_n, y_n)\} \qquad (5\text{-}218) \end{aligned}$$

计算 y_{n+k}． 这是一个 y_{n+k} 的隐式方程．

我们将通过表达式 $e^*_n = y^*_n - y(x_n)$ 来估计

$$e_n = y_n - y(x_n).$$

对于 $m = 0,1,2,\cdots$，我们用关系式

$$f(x_m, y_m) - f(x_m, y(x_m)) = g_m e_m \text{（若 } e_m = 0\text{，则 } g_m = 0\text{），}$$

$$f(x_m, y_m^*) - f(x_m, y(x_m)) = g_m^* e_m^* \text{（若 } e_m^* = 0\text{，则 } g_m^* = 0\text{）}$$

来定义量 g_m 和 g_m^*. 在 (5-216) 和 (5-217) 中令

$$z(x) = y(x_n)$$

并且从对应的关系式 (5-215) 和 (5-218) 中减去这两个式子，我们得到

$$e_{n+k}^* + \alpha_{k-1}^* e_{n+k-1} + \cdots + \alpha_0^* e_n = h\{\beta_{k-1}^* g_{n+k-1} e_{n+k-1} + \cdots + \beta_0^* g_n e_n\} - C_{q+1}^* y^{(q+1)}(x_n) h^{q+1} + O(h^{q+2}),$$

$$\begin{aligned}\alpha_k e_{n+k} &+ \alpha_{k-1} e_{n+k-1} + \cdots + \alpha_0 e_n \\ &= h\{\beta_k g_{n+k}^* e_{n+k}^* + \beta_{k-1} g_{n+k-1} e_{n+k-1} + \cdots + \beta_0 g_n e_n\} \\ &\quad - C_{p+1} y^{(p+1)}(x_n) h^{p+1} + O(h^{p+2}).\end{aligned}$$

消去 e_{n+k}^*，整理后得到

$$\begin{aligned}\alpha_k e_{n+k} &+ \alpha_{k-1} e_{n+k-1} + \cdots + \alpha_0 e_n \\ &= h\{[\beta_{k-1} g_{n+k-1} + \beta_k g_{n+k}^*(-\alpha_{k-1}^* + h\beta_{k-1}^* g_{n+k-1})] e_{n+k-1} \\ &\quad + \cdots + [\beta_0 g_n + \beta_k g_{n+k}^*(-\alpha_0^* + h\beta_0^* g_n)] e_n\} \\ &\quad - C_{p+1} y^{(p+1)}(x_n) h^{p+1} - \beta_k g_{n+k}^* C_{q+1}^* y^{(q+1)}(x_n) h^{q+2} \\ &\quad + O(h^{\min(p+2,q+3)}).\end{aligned} \tag{5-219}$$

我们把引理 5.6 应用于该表达式，令

$$z_m = e_m,\ N = (b-a)/h,\ \beta_{k,m} = 0,$$

$$\beta_{\mu,m} = \beta_\mu g_{m+\mu} + \beta_k g_{m+k}^*(-\alpha_\mu^* + h\beta_\mu^* g_{m+\mu}),$$

$$\mu = 0,1,\cdots,k-1,$$

$\Lambda = Kh^{r+1}$，这里 $r = \min(p, q+1)$，并假设开始值也是 $O(h^{r+1})$. 由 Lipschitz 条件容易得到 $|g_m| \leqslant L$，$|g_m^*| \leqslant L$，从而系数 $\beta_{\mu,m}$ 是有界的. 因此由引理 5.6 我们得出结论：当 $h \to 0$，$x_n = x$ 时，$e_m = O(h^r)$ 对一切 $a \leqslant x \leqslant b$ 的 x 一致地成立. 如果需要，我们可以从 (5-164) 中得到这个定量

估计。我们希望能够很快地看出 e_n 的渐近性态。为了使问题简单化，我们仍假设 $e_\mu = O(h^{r+1})$, $\mu = 0, 1, \cdots, k-1$. 从而，开始误差与真正的离散误差相比较，可以忽略不计。根据 $q+1>p$, $q+1=p$ 或 $q+1<p$，我们必须分三种情形进行讨论。

(i) $q+1>p=r$. 现在我们稍微改变一下 g_m 和 g_m^* 的定义，令

$$g_m = g_m^* = f_y(x_m, y(x_m)).$$

对于这些量 $\bar{e}_n = h^{-r}e_n$，得到代替 (5-182) 的关系式为

$$\begin{aligned}\alpha_k\bar{e}_{n+k} + \cdots + \alpha_0\bar{e}_n = h\{&(\beta_{k-1} - \alpha_{k-1}^*\beta_k)g_{n+k-1}\bar{e}_{n+k-1} \\ &+ \cdots + (\beta_0 - \alpha_0^*\beta_k)g_n\bar{e}_n\} \\ &- hC_{p+1}y^{(p+1)}(x_n) + O(h^2).\end{aligned} \tag{5-220}$$

因为由 $1 + \alpha_{k-1}^* + \cdots + \alpha_0^* = 0$ 知

$$\begin{aligned}(\beta_{k-1} - \alpha_{k-1}^*\beta_k) + \cdots + (\beta_0 - \alpha_0^*\beta_k) &= \beta_{k-1} + \cdots + \beta_0 \\ - (\alpha_0^* + \alpha_1^* + \cdots + \alpha_{k-1}^*)\beta_k &= \beta_{k-1} + \cdots + \beta_0 + \beta_k,\end{aligned}$$

所以 (5-220) 所包含的多步方法是相容的。

由 $y^{(p+1)}(x)$ 的可微性及误差常数

$$C = \frac{C_{p+1}}{\beta_0 + \beta_1 + \cdots + \beta_k}$$

的定义，(5-220) 中第三行为

$$\begin{aligned}-h\{&(\beta_{k-1} - \alpha_{k-1}^*\beta_k)Cy^{(p+1)}(x_{n+k-1}) + \cdots \\ &+ (\beta_0 - \alpha_0^*\beta_k)Cy^{(p+1)}(x_n)\} + O(h^2).\end{aligned}$$

应用定理 5.11 得

$$\bar{e}_n = e(x_n) + O(h), \tag{5-221}$$

这里

$$\begin{aligned}&e'(x) = g(x)e(x) - Cy^{(p+1)}(x), \\ &e(a) = 0.\end{aligned} \tag{5-222}$$

于是，如果预估公式的阶数至少是校正公式的阶数时，那

么仅用一次校正和无穷次校正的离散误差的渐近性态是相同的。

(ii) $q+1=p=r$. 对 $\bar{e}_n=h^{-r}e_n$，我们得到了关系式 (5-220) 中的第三行以

$$-h[C_{p+1}y^{(p+1)}(x_n)+\beta_k g(x_n)C_p^* y^{(p)}(x_n)]+O(h^2)$$

来代替. 根据定理 11，得到

$$\bar{e}_n=e(x_n)+O(h), \tag{5-223}$$

这里

$$e'(x)=g(x)e(x)-Cy^{(p+1)}(x)-\frac{\beta_k C_p^*}{\beta_k+\cdots+\beta_0}g(x)y^{(p)}(x),$$

$$e(a)=0. \tag{5-224}$$

因此，如果校正公式的阶超过预估公式的阶为 1 时，仅用一次校正离散误差的阶不变；但是，在伸缩误差函数 $e(x)$ 的定义中出现了一个外加的项. 在下面的特殊情形：

$$y_{n+1}^*-y_n=hf_n,\ \ y_{n+1}-y_n=\frac{1}{2}h(f_{n+1}+f_n),$$

我们得到一个与 §2.2-5 中简化的 Runge-Kutta 方法 [方程 (2-10b)] 有关的结果，这是用完全不同的方法得到的[参看方程 (2-35)，$\alpha=1/2$].

(iii) $r=q+1<p$. 我们再一次得到对 $\bar{e}_n=h^{-r}e_n$ 的关系式 (5-220)，而现在最后一行由

$$-h\beta_{q+1}^* C_{q+1}^* g(x_n)y^{(p+1)}(x_n)+O(h^2)$$

来代替.

再由定理 5.11，我们得到 $\bar{e}_n=e(x_n)+O(h)$，这里

$$e'(x)=g(x)e(x)-\frac{\beta_k C_{q+1}^*}{\beta_0+\cdots+\beta_k}g(x)y^{(q+1)}(x),$$

$$e(a)=0. \tag{5-225}$$

所以，如果校正公式的阶超过预估公式的阶 >1，那么仅

用一次校正时，离散误差的阶是减少的，并且改变了伸缩误差函数的形式。

从某种程度上说，这部分结果完全平行于§3.2-4中所讨论的单步方法用求积公式得到的结果．仅用一次校正的算法在实际中被广泛采用．

5.4. 多步方法积分的舍入误差

5.4-1. 一个先验界．我们用实际计算满足方程

$$\alpha_k\tilde{y}_{n+k}+\alpha_{k-1}\tilde{y}_{n+k-1}+\cdots+\alpha_0\tilde{y}_n=h\{\beta_k f(x_{n+k},\tilde{y}_{n+k})+\cdots+\beta_0 f(x_n,\tilde{y}_n)\}+\varepsilon_{n+k},\quad n=0,1,2,\cdots \tag{5-226}$$

的量 $\tilde{y}_n$ 来代替精确满足差分方程(5-58)的量．

量 ε_n 便称为局部舍入误差．在§5.4.中所要讨论的问题是这些局部误差对累积舍入误差 $r_n=\tilde{y}_n-y_n$ 的影响．

不作任何关于舍入误差性质的推测，在这一节中，我们将在单独假设 $|\varepsilon_{n+k}|\leqslant\varepsilon(n=0,1,2,\cdots)$ 的前提下，导出 r_n 的先验估计，这里 ε 是一个常数．由(5-226)减去对应的方程(5-58)，并令

$$g_m=r_m^{-1}[f(x_m,\tilde{y}_m)-f(x_m,y_m)],\quad 若\ r_m\neq 0,$$
$$g_m=0,\quad 若\ r_m=0.$$

于是恒有 $|g_m|\leqslant L$，L 为 Lipschitz 常数，我们得到

$$\alpha_k r_{n+k}+\alpha_{k-1}r_{n+k-1}+\cdots+\alpha_0 r_n=h\{\beta_k g_{n+k}r_{n+k}+\cdots+\beta_0 g_n r_n\}+\varepsilon_{n+k}. \tag{5-227}$$

把引理5.6应用到该关系式，取

$$z_m=r_m,\quad \Lambda=\varepsilon,\quad N=(b-a)/h,$$

以及(因为 $r_0=r_1=\cdots=r_{k-1}=0$) $Z=0$，得到

$$r_n\leqslant\varepsilon h^{-1}(x_n-a)\Gamma^*\exp[(x_n-a)\Gamma^*BL], \tag{5-228}$$

其中这些常数象在定理5.11那样确定．

关系式 (5-228) 可能以一个大的幅度过高地估计了实际舍入误差。它所包含的实质而又对后面的推导过程有着重要意义的结果是 $r_n = O(\varepsilon h^{-1})$; 根据后面 (§6.3-2) 对二阶方程所要推导的结果,这是有价值的。

5.4-2 一个后验界。现在我们假定 $Nh^{p+1} \leqslant \varepsilon \leqslant Kh^2$, 其中 N 和 K 与 h 无关。于是由 (5-228), 我们有 $r_n = O(h)$. 假设 $f_{yy}(x,y)$ 在 $y = y(x)$ 的一个邻域内存在且连续,那么如果 h 充分小,我们可以写

$$f(x_m, \tilde{y}_m) - f(x_m, y_m) = g(x_m) r_m + \theta_m K_2 \varepsilon, \quad |\theta_m| \leqslant 1,$$

其中 $g(x) = f_y(x, y(x))$. 从而方程 (5-227) 可以替换成

$$\alpha_k r_{n+k} + \cdots + \alpha_0 r_n = h\{\beta_k g_{n+k} r_{n+k} + \cdots + \beta_0 g_n r_n\} + \varepsilon_{n+k} + \theta_n K_3 h \varepsilon, \qquad (5\text{-}229)$$

这里 $g_m = g(x_m)$. 我们可以相应地写成

$$r_n = r_n^{(1)} + r_n^{(2)},$$

其中 $\{r_n^{(1)}\}$ 是 $\theta_n = 0$ 时 (5-229) 的解,而 $\{r_n^{(2)}\}$ 是 $\varepsilon_{n+k} = 0$ 时的解,这两个解都满足 $r_\mu^{(i)} = 0$, $\mu = 0, 1, \cdots, k-1$. 我们可以称 $r_n^{(1)}$ 为主要误差,而称 $r_n^{(2)}$ 为次要误差。我们注意到次要误差仅仅是由于微分方程的非线性性所引起的。由引理 5.6 知,次要误差为 $O(\varepsilon)$, 而主要误差必须是 $O(\varepsilon h^{-1})$. 因此我们可以假定当 $h \to 0$ 时 r_n 的性态是受 $r_n^{(1)}$ 的性态支配的。下面我们就主要误差 $r_n^{(1)}$ 来考虑。

根据定理 5.2, 主要误差可表示成

$$r_n^{(1)} = \sum_{l=k}^{n} \varepsilon_l d_{nl}, \qquad (5\text{-}230)$$

而对于 $l = k, k+1, \cdots, \{d_{nl}\}$ 是差分方程

$$\alpha_k d_{n+k,l} + \alpha_{k-1} d_{n+k-1,l} + \cdots + \alpha_0 d_{n,l} = h\{\beta_k g_{n+k} d_{n+k,l} + \cdots + \beta_0 g_n d_{n,l}\}, \quad (5\text{-}231)$$

的解,假定初值为

$$d_{n,l}=\begin{cases}0, & n<l,\\ \dfrac{1}{\alpha_k-h\beta_k g_l}, & n=l.\end{cases}\qquad (5-232)$$

把 (5-231) 和 (5-232) 式与 (5-184) 和 (5-183) 进行比较，我们看到，$d_{n,l}$ 的初值问题与 §5.3-5 中所讨论的量 e_n^{-H} 的值是相同的。如果 x_0 用 x_{l-k+1} 代替，并对初值函数选取特殊函数

$$\delta_\mu(h)=0,\mu=0,\quad 1,\cdots,k-2,$$

$$\delta_\mu(h)=\frac{h^r}{\alpha_k-h\beta_k g_l},\quad \mu=k-1.$$

那么，我们可以应用 (5-231) 式。注意到

$$\rho_\mu(\zeta)=\rho(\zeta)/(\zeta-\zeta_\mu)=\alpha_k\zeta^{k-1}+\cdots,$$

我们有 $\alpha_{\mu,k-1}=\alpha_k$，从而 $\Delta_\mu=1$. 于是有下面的结果：

$$d_{n,l}=\sum_{\mu=1}^{m}\frac{1}{\rho'(\zeta_\mu)}\exp[i(n-l+k-1)\varphi_\mu]d_{l,\mu}(x_n)+O(h),\qquad (5-233)$$

其中函数 $d_{l,\mu}(x)$ 由类似于 (5-197) 的初值问题

$$d'_{l,\mu}(x)=\lambda_\mu g(x)d_{l,\mu}(x),$$

$$d_{l,\mu}(x_l)=1,\quad \mu=1,2,\cdots,m\qquad (5-234)$$

所确定(在这点上增长参数 λ_μ 是舍入误差传播的决定影响)。容易证明，对于 $x\geqslant x_l$

$$d_{l,\mu}(x)=\frac{e_\mu(x)}{e_\mu(x_l)},\quad \mu=1,2,\cdots,m,$$

其中函数 $e_\mu(x)$ 由 (5-197) 确定. 代入 (5-230)，得到

$$r_n^{(1)}=\sum_{l=k}^{n}\varepsilon_l\left\{\sum_{\mu=1}^{m}\frac{\exp[i(n-l+k-1)\varphi_\mu]}{\rho'(\zeta_\mu)}\,\frac{e_\mu(x_n)}{e_\mu(x_l)}+O(h)\right\}.\qquad (5-235)$$

作为该公式的第一个应用，我们将在假设

$$|\varepsilon_l| \leqslant p(x_l)\varepsilon \qquad (5-236)$$

下得到 $r_n^{(1)}$ 的渐近界，这里 ε 不依赖于 x，而 $p(x)$ 是已知的 x 的连续(或分段连续)函数．重新排列 (5-235) 中求和的次序，我们得到

$$|r_n^{(1)}| \leqslant \frac{\varepsilon}{h} \sum_{\mu=1}^{m} \frac{1}{|\rho'(\zeta_\mu)|} m_{\mu,n},$$

其中

$$m_{\mu,n} = |e_\mu(x_n)| h \sum_{l=k}^{n} \{p(x_l)|e_\mu(x_l)|^{-1} + O(h)\}. \qquad (5-237)$$

在这一点上，插进下面的简单引理，对以后是方便的，因为它将被反复用到。

引理 5.8. 设 $f(x)$ 在 $[a,b]$ 上连续，并且连续可微，对于 $x_n \in [a,b]$，令

$$\Sigma_n = h \sum_{p=0}^{n} \{f(x_p) + h\theta_p K\}, \qquad (5-238)$$

其中 $|\theta_p| \leqslant 1$，而 K 是一个与 h 或 n 无关的常数． 因此存在一个常数 K_1，使得

$$\left|\Sigma_n - \int_a^{x_n} f(t)dt\right| \leqslant hK_1. \qquad (5-239)$$

这是定理1.4的一个特殊情形[$f(x,y)$ 不依赖于 y]．直观上和 Σ_n 可看成近似于一个定积分的 Riemann 和。

函数 $f(x) = p(x)|e_\mu(x)|^{-1}$ 满足该引理的条件，因为当 $x \in [a,b]$ 时 $p(x)$ 和 $e_\mu(x)$ 是可微的，并且 $e_\mu(x) \neq 0$．因此我们可以把这个引理应用于 (5-237) 中出现的和上，从而得到

$$m_{\mu,n} = m_\mu(x_n) + O(h),$$

其中

$$m_\mu(x) = |e_\mu(x)| \int_a^x p(t) |e_\mu(t)|^{-1} dt. \qquad (5-240)$$

因此我们得到

定理 5.14. 如果局部舍入误差 ε_l 满足 (5-236)，而

$$\varepsilon = O(h^2),$$

那么

$$|r_n^{(1)}| = \frac{\varepsilon}{h} \sum_{\mu=1}^{m} \{|\rho'(\zeta_\mu)|^{-1} m_\mu(x_n) + O(h)\}, \qquad (5-241)$$

其中函数 $m_\mu(x)$ 由 (5-240) 所确定。

和单步方法的情形一样，我们把函数 $m_\mu(x)$ 表现为某个初值问题的解[参看(2-77)]. 暂时设 $e_\mu(x) = u(x) + iv(x)$ 及 $\lambda_\mu = \xi + i\eta$，则由 (5-197) 式分开实部与虚部，我们有

$$u' = (\xi u - \eta v)g, \quad v' = (\xi v + \eta u)g.$$

于是，如果 $|e_\mu| = (u^2 + v^2)^{1/2} = r$，则我们得到

$$rr' = uu' + vv' = (\xi u^2 - \eta uv + \xi v^2 + \eta uv)g = \xi r^2 g,$$

因此

$$|e_\mu|' = \mathrm{Re}\, \lambda_\mu g |e_\mu|.$$

微分 (5-240) 式，我们容易得到

$$\begin{cases} m_\mu'(x) = \mathrm{Re}\, \lambda_\mu g(x) m_\mu(x) + p(x), \\ m_\mu(a) = 0. \end{cases} \qquad (5-242)$$

于是，就我们所涉及到的函数 $m_\mu(x)$ 的增长来说，只有增长参数 λ_μ 的实部是重要的。

5.4-3. 统计估计. 前面已经指出过，真正令人满意的舍入误差理论只有在统计的基础上才能达到。尤其对于多步方法更是如此。特别是，统计方法指出了条件稳定的实际性质。将要指出由实部为负的增长参数引起的条件稳定是不受开始误差限制的。以定量可预估的方法来看，条件稳定是由于舍

入误差引起的。

和 §2.3-4 中一样，我们假定局部舍入误差 ε_l 都是独立的随机变量，其均值和方差满足

$$|E(\varepsilon_l)| \leqslant \mu\, p(x_l), \tag{5-243}$$

$$\operatorname{var}(\varepsilon_l) = \sigma^2 q(x_l), \tag{5-244}$$

其中 $p(x)$ 和 $q(x)$ 已知为 $[a,b]$ 上非负的分段光滑函数。假定量 μ 和 σ^2 与 x 无关。对于各种算术系统来说，关于这些函数和量的适当假设在 §5.4-4 中讨论。那时足以指出对隐式线性多步方法来说 $\mu=0$ 的假设在许多情况下是不现实的。

由 (5-235)，利用 (1-94)，我们有

$$E(r_n^{(1)}) = \sum_{l=k}^{n} E(\varepsilon_l)\left\{\sum_{\mu=1}^{m} \frac{\exp[i(n-l+k-1)\varphi_\mu]}{\rho'(\zeta_\mu)} \times \frac{e_\mu(x_n)}{e_\mu(x_l)} + O(h)\right\}.$$

利用 (5-243)，象 (5-241) 的推导那样，于是就有

$$|E(r_n^{(1)})| \leqslant \frac{\mu}{h}\left\{\sum_{\mu=1}^{m} |\rho'(\zeta_\mu)|^{-1} m_\mu(x_n) + O(h)\right\}, \tag{5-245}$$

其中函数 $m_\mu(x)$ 由 (5-242) 确定。

$\operatorname{var}(r_n^{(1)})$ 的计算较复杂。为了说明这个方法，我们首先考虑单个方程 $y' = Ay$（A 为实常数），假定 $q(x)=1$。由 §5.3-2 我们知道，此时 $e_\mu(x) = e^{\lambda_\mu A x}$。从而方程 (5-235) 变成

$$r_n^{(1)} = \sum_{l=k}^{n} \varepsilon_l d_{nl},$$

其中

$$d_{nl} = \sum_{\mu=1}^{m} \frac{1}{\rho'(\zeta_\mu)} \exp[i(n-l+k-1)\varphi_\mu] \times \exp[\lambda_\mu A(x_n - x_l)] + O(h). \tag{5-246}$$

利用 (5-95)，我们有

$$\operatorname{var}(r_n^{(1)})=\frac{\sigma^2}{h}v_n,\quad v_n=h\sum_{l=k}^{m}d_{nl}^2.$$

由它们的性质可知，系数 d_{nl} 是实数．因此，我们可以令

$$d_{nl}^2=d_{nl}\bar{d}_{nl},$$

这里"—"表示共轭复数．由 (5-246)，我们得到

$$d_{nl}\bar{d}_{nl}=\Big(\sum_{\mu=1}^{m}\frac{1}{\rho'(\zeta_\mu)}\exp[i(n-l+k-1)\varphi_\mu]$$
$$\times\exp[\lambda_\mu A(x_n-x_l)]\Big)$$
$$\times\Big(\sum_{\nu=1}^{m}\frac{1}{\overline{\rho'(\zeta_\nu)}}\exp[-i(n-l+k-1)\varphi_\nu]$$
$$\times\exp[\bar{\lambda}_\nu A(x_n-x_l)]\Big)+O(h).$$

对于 $\mu=\nu$ 选出交叉项，这个式子可写为

$$d_{nl}\bar{d}_{nl}=\sum_{\mu=1}^{m}|\rho'(\zeta_\mu)|^{-2}\exp[2\mathrm{Re}\,\lambda_\mu A(x_n-x_l)]$$
$$+\sum_{\substack{\mu,\nu=1\\ \mu\neq\nu}}^{m}C_{\mu\nu}\exp[i(n-l+k-1)\delta_{\mu\nu}]\exp[(\lambda_\mu+\bar{\lambda}_\nu)$$
$$\times A(x_n-x_l)]+O(h),$$

其中

$$C_{\mu\nu}=[\rho'(\zeta_\mu)\overline{\rho'(\zeta_\mu)}]^{-1},\quad \delta_{\mu\nu}=\varphi_\mu-\varphi_\nu.$$

求 v_n 所需的和，可以通过计算几何级数来完成，其结果是(设 $A\neq 0$)

$$h\sum_{l=k}^{n}\exp[2\mathrm{Re}\lambda_\mu A(x_\mu-x_l)]=h\frac{\exp[2\mathrm{Re}\,\lambda_\mu Ah(n-k+1)]-1}{\exp(2\mathrm{Re}\,\lambda_\mu Ah)-1}$$
$$=\frac{\exp[2\mathrm{Re}\,\lambda_\mu Ax_n]-1}{2\mathrm{Re}\,\lambda_\mu A}+O(h).\qquad(5\text{-}247)$$

若 $\mu \neq \nu$，则

$$h \sum_{l=k}^{n} \exp[i(n-l+k-1)\delta_{\mu\nu}]\exp[(\lambda_\mu+\lambda_\nu)A(x_n-x_l)]$$

$$= h\exp[i(k-1)\delta_{\mu\nu}] \sum_{p=0}^{n-k} \exp\{[(\lambda_\mu+\bar{\lambda}_\nu)Ah+i\delta_{\mu\nu}]p\}$$

$$= h\exp[i(k-1)\delta_{\mu\nu}]$$

$$\times \frac{\exp\{[(\lambda_\mu+\bar{\lambda}_\mu)Ah+i\delta_{\mu\nu}](n-k+1)\}-1}{\exp[(\lambda_\mu+\lambda_\mu)Ah+i\delta_{\mu\nu}]-1}. \quad (5\text{-}248)$$

因为 $\delta_{\mu\nu} \neq 0 \pmod{2\pi}$（对于一个稳定的方法，$\zeta_1, \zeta_2, \cdots, \zeta_m$ 都是单根)，最后函数中的分母当 $h \to 0$ 时趋向于

$$\exp(i\delta_{\mu\nu}) - 1 \neq 0.$$

因此最后的表达式可以写成

$$h\frac{\exp(in\delta_{\mu\nu})\exp[(\lambda_\mu+\bar{\lambda}_\nu)Ax_n] - \exp[i(k-1)\delta_{\mu\nu}]}{\exp(i\delta_{\mu\nu})-1}$$

$$+ O(h^2).$$

由此可见，$\mu \neq \nu$ 所给出的交叉积对 v_n 的贡献为 $O(h)$. 所以

$$v_n = \sum_{\mu=1}^{m} |\rho'(\zeta_\mu)|^{-2} \frac{\exp(2\mathrm{Re}\,\lambda_\mu Ax_n)-1}{2\mathrm{Re}\,\lambda_\mu A} + O(h).$$

$$(5\text{-}249)$$

在 (5-249) 中出现的函数

$$v_\mu(x) = \frac{\exp(2\mathrm{Re}\,\lambda_\mu Ax)-1}{2\mathrm{Re}\,\lambda_\mu A}$$

可以表征为

$$\begin{aligned} v'_\mu(x) &= 2\mathrm{Re}\,\lambda_\mu A v_\mu(x) + 1 \\ v_\mu(0) &= 0, \quad \mu = 1, 2, \cdots, m \end{aligned} \quad (5\text{-}250)$$

的解.

现在我们回到一般情形. 这里的几何级数求和的技巧必

数由更为广泛的应用方法来代替. 方程 (5-230) 成立，取

$$d_{nl}=\sum_{\mu=1}^{m}\frac{1}{\rho'(\zeta_\mu)}\exp[i(n-l+k-1)\varphi_\mu]e_\mu(x_n)[e_\mu(x_l)]^{-1}+O(h),$$

并且有

$$\operatorname{var}(r_n^{(1)})=\frac{\sigma^2}{h}v_n;\quad v_n=h\sum_{l=k}^{n}q_l d_{nl}^2.$$

这里 $q_l=q(x_l)$. 象上面特殊情形所进行的那样，我们得到

$$d_{nl}\bar{d}_{nl}=\sum_{\mu=1}^{m}C_{\mu\mu}d_{nl\mu\mu}+\sum_{\substack{\mu,\nu=1\\ \mu\neq\nu}}^{m}C_{\mu\nu}\exp[i(n-l+k-1)\delta_{\mu\nu}]d_{nl\mu\nu}+O(h),\tag{5-251}$$

这里当 $\mu,\nu=1,\cdots,m$ 时

$$\begin{aligned}C_{\mu\nu}&=[\rho'(\zeta_\mu)\overline{\rho'(\zeta_\nu)}]^{-1},\\ \delta_{\mu\nu}&=\varphi_\mu-\varphi_\nu,\\ d_{nl\mu\nu}&=e_\mu(x_n)[e_\mu(x_l)\overline{e_\nu(x_l)}]^{-1}\overline{e_\nu(x_n)}.\end{aligned}\tag{5-252}$$

显然

$$v_n=\sum_{\mu=1}^{m}C_{\mu\mu}v_{n\mu\mu}+\sum_{\substack{\mu,\nu=1\\ \mu\neq\nu}}^{n}C_{\mu\nu}v_{n\mu\nu}+O(h),$$

其中

$$v_{n\mu\nu}=h\sum_{l=k}^{n}q_l d_{nl\mu\nu}\exp[i(n-l+k-1)\delta_{\mu\nu}].\tag{5-253}$$

我们先来求 $v_{n\mu\mu}$ 的值. 考虑和 (5-253) 作为 Riemann 积分的近似值，由引理 5.8，我们有

$$v_{n\mu\mu}=v_\mu(x_n)+O(h),$$

其中

$$v_\mu(x)=e_\mu(x)\overline{e_\mu(x)}\int_a^x q(t)[e_\mu(t)\overline{e_\mu(t)}]^{-1}dt.$$

利用(5-197)式，我们容易验证函数 $v_\mu(x)$ 满足微分方程

$$\begin{aligned}&v_\mu'(x)=2\mathrm{Re}\,\lambda_\mu g(x)v_\mu(x)+q(x),\\&v_\mu(a)=0,\ \mu=1,2,\cdots,m.\end{aligned}\qquad(5\text{-}254)$$

这是(5-250)式的推广.

现在我们要证明，当 $\mu\neq\nu$ 时，$v_{n\mu\nu}=O(h)$。这可以由下面的引理来完成。

引理 5.9. 设函数 $f(x)$ 在 $a\leqslant x\leqslant b$ 上连续且连续可微，并设 $\delta\not\equiv 0(\mathrm{mod}\ 2\pi)$，则存在一个常数 C，使得

$$\left|\sum_{p=0}^{n}e^{ip\delta}f(x_p)\right|\leqslant C\qquad(5\text{-}255)$$

对所有 $h>0$，$x_n\in[a,b]$ 都成立.

证明是采用求部分和的方法。令

$$S_p=\sum_{l=0}^{p}e^{il\delta}=\frac{e^{i(p+1)\delta}-1}{e^{i\delta}-1},$$

我们有

$$\sum_{p=0}^{n}e^{ip\delta}f(x_p)=f(x_0)+\sum_{p=1}^{n}(S_p-S_{p-1})f(x_p)$$

$$=\sum_{p=0}^{n-1}S_p[f(x_p)-f(x_{p+1})]+S_nf(x_n).$$

由中值定理，对某个 $\xi\in[x_p,x_{p+1}]$，有

$$f(x_p)-f(x_{p+1})=-hf'(\xi).$$

利用

$$|S_p|\leqslant\frac{2}{|e^{i\delta}-1|}$$

以及 $n\leqslant(b-a)/h$，取

$$C=\frac{2}{|e^{i\delta}-1|}\left\{(b-a)\max_{x\in[a,b]}|f'(x)|+\max_{x\in[a,b]}|f(x)|\right\},$$
$$(5\text{-}256)$$

我们便得 (5-255). 把引理 5.8 应用于函数

$$f(t)=q(t)|e_\mu(t)\overline{e_\nu(t)}|^{-1}.$$

它满足连续且可微的条件,因为 $e_\mu(t)\neq 0$——令 $\delta=\delta_{\mu\nu}$,当 $\mu\neq\nu$ 时便证明了 $v_{n\mu\nu}=O(h)$. 于是我们可以叙述下面的主要结果:

定理 5.15. 如果局部舍入误差是满足(5-243)和(5-244)的独立随机变量,那么累积舍入误差的主要分量是一个随机变量,其均值满足 (5-245) 而方差为

$$\operatorname{var}(r_n^{(1)})=\frac{\sigma^2}{h}\left\{\sum_{\mu=1}^{m}|\rho'(\zeta_\mu)|^{-2}v_\mu(x_n)+O(h)\right\},$$
$$x_n\in[a,b], \qquad (5\text{-}257)$$

函数 $v_\mu(x)$由 (5-254) 所确定.

对于 $a\leqslant x\leqslant b$, 如果 $q(x)>0$, 那么容易证明条件(1-96) 是满足的, 因此当 $n\to\infty$时, $r_n^{(1)}$ 的分布服从中心极限定理.

应当注意, 表达式 (5-257) 中函数 $v_\mu(x)$ 的相关强度仅仅依赖于多项式 $\rho(\zeta)$, 因此也就依赖于作为基础的积分公式. 要从 (5-257) 中以特殊的运算方法消去不希望有的项, 是不可能的.

5.4-4. *局部舍入误差.* 假定 $\alpha_k=1$, 从而新的值 y_{n+k} 可以由公式

$$y_{n+k}=-\alpha_{k-1}y_{n+k-1}-\cdots-\alpha_0y_n$$
$$+h\{\beta_kf_{n+k}+\cdots+\beta_0f_n\} \qquad (5\text{-}258)$$

计算得到 (因为一开始已经给出假设 $\alpha_k\neq 0$, 所以是可以作这个假定的,但似乎保持公式的对称形式较好). 公式的右边是变量 x_n, y_n, y_{n+1}, y_{n+k-1} 和 h 的函数 F. 当 $\beta_k=0$ 时这是明

能的，而当 $\beta_k \neq 0$ 时，只要 h 充分小，由定理5.4 它是成立的。设 x_n 和 h 是精确数，则局部舍入误差是函数 F 的计算值与数学值之差，两者都在点 $\tilde{y},\cdots,\tilde{y}_{n+k-1}$ 处求值．局部舍入误差的研究是复杂的，每个特殊的方法和运算过程都必须根据它自己的特点来下结论．我们仅限于作一些定性的一般性的评论。

(i) 定点运算． 我们首先讨论 $\beta_k = 0$ 的情形． 如果用单倍位精确度运算，并且表达式 $\Phi = \beta_{k-1}f_{n+k-1} + \cdots + \beta_0 f_n$ 第一次求值，那么 Φ 的固有误差阶为 hu(u = 基本单位)并且与 $h\Phi$ 形成的引入误差（该误差为 u 的阶）相比是可以忽略的．如果系数 α_μ 中有些不是整数，那么乘积 $\alpha_\mu y_{n+\mu}$ 产生附加误差．这些误差具有同样的数量级和像引入误差一样的变化范围．如果 $\beta_k \neq 0$，则表达式

$$C = -\alpha_{k-1}y_{n+k-1} - \cdots \alpha_0 y_n + h\{\beta_{k-1}f_{n+k-1} + \cdots + \beta_0 f_n\}$$

通常是只求一次值，并且有上述的误差．附加的误差是由包含函数 $h\beta_k f(x_{n+k}, y_{n+k})$ 的反复求值的迭代过程所引起的．由逐次迭代所产生的误差不积累，而相当于一个引入误差大小的附加误差．该误差可以称为迭代误差．经验表明，由于迭代过程的极限 y_{n+k} 是从同一侧反复逼近的，故迭代误差常常是有偏差的．如果使用部分的双倍位精确度，那么引入误差为零． 因此局部误差包括由截断 f 的变量 $y_{n+\mu}$ 所引起的固有误差，以及可能产生的迭代误差，阶全部是 hu． 如果系数 α_μ 中有些不是整数，那么为了避免大小为 u 的误差，乘积 $\alpha_\mu y_{n+\mu}$ 必须以双倍位精确度形成．

(ii) 浮点运算． 固有的和引入的误差现在都是 hfu 的阶，其中 u 表示尾数的基本单位．另一方面，形成 C 所产生的误差的阶为 uy。

这个误差可以称为主导误差．主导误差的数量级的另一

个误差在隐式方法中当最后的校正量 $h\beta_k f_{n+k}$ 加到C 上去时产生。

至于误差中统计的假设，最妥善的假设是包括假设局部误差均匀地分布在它的极值之间．理论上较好的结果通常可以通过把局部误差细分到它的各个分量和通过假定每个分量是一个均匀地分布在它的最大和最小可能值之间的独立随机变量来得到。

5.4-5. **数值例子**．我们用下面的三个多步方法来积分初值问题

$$y' = -16xy,$$
$$y(-0.75) = y_{0,q} = y_{0,0}(1 + q\Delta), \quad q = 0, \cdots, Q-1. \tag{5-259}$$

（在 §2.3-7 中已经讨论过）对上面得到的统计结果作了试验。

（I）二阶 Adams-Bashforth 方法

$$y_{n+2} - y_{n+1} = h\left\{\frac{3}{2} f_{n+1} - \frac{1}{2} f_n\right\};$$

（II）二阶 Nyström 方法（"中点公式"）

$$y_{n+2} - y_n = 2hf_{n+1};$$

（III）四阶 Milne 方法

$$y_{n+2} - y_n = h\left\{\frac{1}{3} f_{n+2} + \frac{4}{3} f_{n+1} + \frac{1}{3} f_n\right\}.$$

这三种方法都使用下面的数据：

$$h = 2^{-6},$$
$$y_{0,0} = 0.0022159242 \cdot 2^{-15},$$
$$y_{-1,0} = 0.0018334831 \cdot 2^{-15} \text{（精确值）},$$
$$\Delta = \frac{1}{3} \cdot 2^{-8},$$
$$Q = 500,$$
$$u = 2^{-36}.$$

精确值 $y_{n,0}$ 是用较高精确度数值经计算得到。这三种情形讨论的结果如下。

(I) 因为 $m=1, \rho'(1)=1$，故由 (5-257) 所表示的模型与单步法所确定的模型是相同的。和适合定点运算时那样假定 $q(x)=1$，则 $\nu_1(x)$ 满足

$$\nu_1'=1-32x\nu_1, \quad \nu_1(-0.75)=0. \qquad (5\text{-}260)$$

通过数值积分得到的 $\nu_1(x)$ 的数值在表 5.16-I 中给出。

表 5.16-I

x		−0.50	−0.25	0	0.25	0.50	0.75
$\nu_1(x)$		6.5	132.7	361.1	133.0	6.7	0.1
$u^{-1}E(r_n)$	试验	1.1	5.1	8.8	4.7	0.8	0.0
	预估	0	0	0	0	0	0
$u^{-2}\text{var}(r_n)$	试验	47.7	972.1	2586.5	931.0	44.1	0.5
	预估	34.9	707.7	1925.9	709.3	35.7	0.5

设 $\mu=0, \sigma^2=\frac{1}{12}u^2$，所得到的预估的和试验的均值与方差在表 5.16-I 中给出。方差的试验值以一个近似定比而超过预估值。这个偏差看来是由于忽略了次要的和固有的误差产生的。

(II) 与中点公式的根 $\zeta_1=1$ 和 $\zeta_2=-1$ 有关的增长参数 λ_1 和 λ_2 在 §5.3-1 中已经确定为 $\lambda_1=1$，$\lambda_2=-1$。由于 $\rho'(1)=2, \rho'(-1)=-2$，故我们现在要求近似地有

$$\text{var}(r_n^{(1)})=\frac{\sigma^2}{4h}(\nu_1(x_n)+\nu_2(x_n)),$$

其中 $\nu_1(x)$ 由 (5-260) 给定，而 $\nu_2(x)$ 满足

$$\nu_2'(x)=1+32x\nu_2, \quad \nu_2(-0.75)=0. \quad (5\text{-}261)$$

(5-261)数值积分得到的值在表 5.16-II 中给出。

表 5.16-II

x	-0.50	-0.25	0	0.25	0.50	0.75
$\nu_2(x)$	0.1	0.1	0.2	1.1	24.1	3590.5
$u^{-1}E(r_n)$ 试验	0.4	2.6	4.3	2.3	0.5	2.7
$u^{-1}E(r_n)$ 预估	0	0	0	0	0	0
$u^{-2}\mathrm{var}(r_n)$ 试验	21.6	395.5	981.5	330.2	55.3	6335.5
$u^{-2}\mathrm{var}(r_n)$ 预估	16.6	357.0	963.4	357.6	82.2	9574.0

表 5.16-II 中给出的预估值是用 $\mu=0$，$\sigma^2=\frac{1}{6}u^2$ 计算的，并给出了数值结果。对于 $x\geqslant 0.50$，由于 $\nu_2(x)$ 的增长导致 $\mathrm{var}(r_n)$ 的急剧增长，与解的指数递减形成鲜明的对照。

(III) Milne-Simpson 过程的增长参数确定为

$$\lambda_1=1,\ \lambda_2=-\frac{1}{3}.$$

又因为 $\rho'(1)=-\rho'(-1)=2$，我们要求

$$\mathrm{var}(r_n^{(1)})=\frac{\sigma^2}{4h}(\nu_1(x)+\nu_2(x)),$$

表 5.16-III

x	-0.50	-0.25	0	0.25	0.50	0.75
$\nu_2(x)$	0.1	0.2	0.4	0.8	2.7	15.2
$u^{-1}E(r_n)$ 试验	18.3	87.0	141.6	84.7	20.1	2.6
$u^{-1}E(r_n)$ 预估	—	—	—	—	—	—
$u^{-2}\mathrm{var}$ 试验	86.9	1518.9	3751.4	1270.7	79.1	117.2
$u^{-2}\mathrm{var}$ 预估	70.4	1417.6	3856.0	1427.2	100.2	163.6

其中 $\nu_1(x)$ 由 (5-260) 所确定，而 $\nu_2(x)$ 由

$$\nu_2'=1+\frac{1}{3}\cdot 32x\nu_2,\quad \nu_2(-0.75)=0 \qquad (5\text{-}262)$$

所确定。$\nu_2(x)$ 的数值在表 5.16-III 中给出。

—— · ——预估的标准偏差

——计算的标准偏差

……计算的平均值

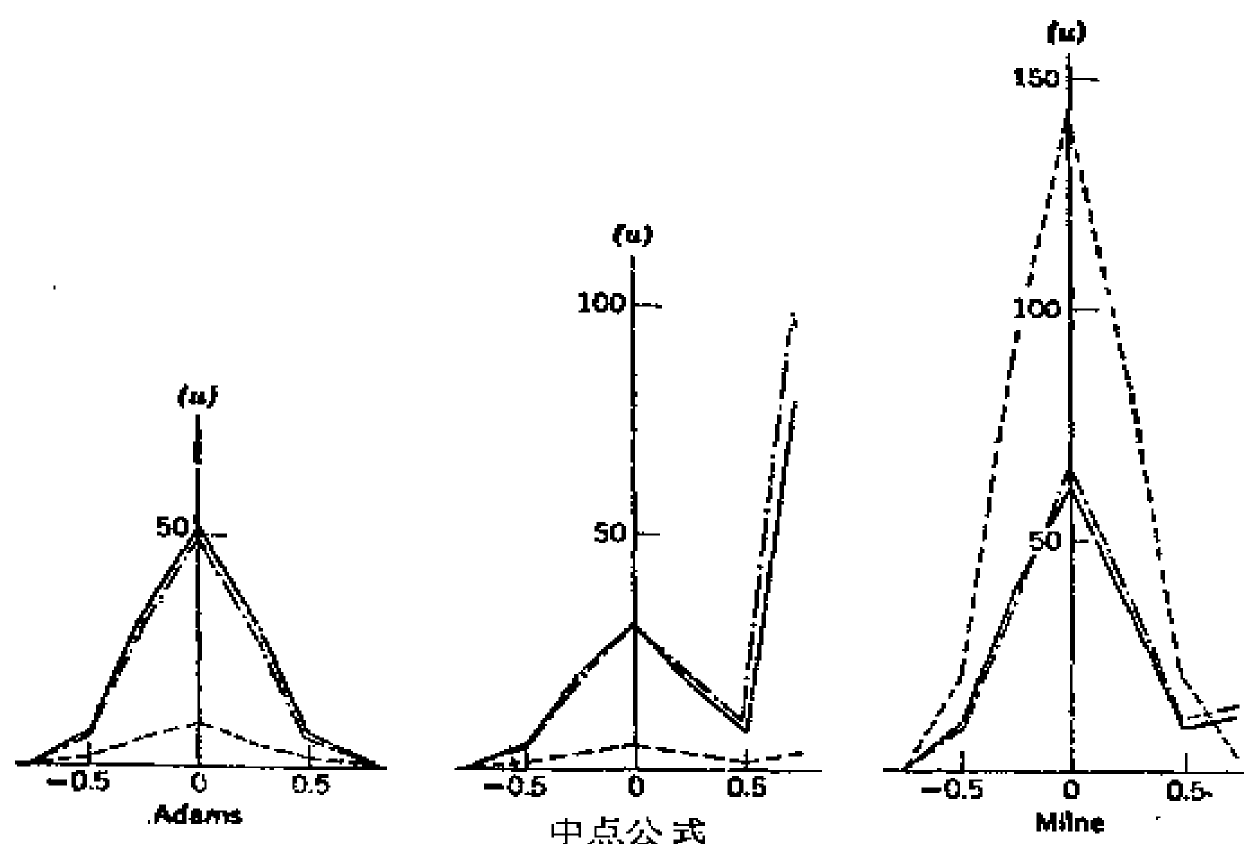


图 5.3

$x \geqslant 0$ 同样有增长，虽然不象中点公式那样明显。在表5.16-III 中并没有试图预估 $E(r_n)$；而假设 $\mu = 0$ 显然是不现实的。对于局部误差的方差，根据现在出现迭代误差的事实我们假定 $\sigma^2 = \frac{2}{3}u^2$。（使用中点公式获得首次预估值）还应当注意到 $\mathrm{var}(r_n)$ 值的最后向上摆动的不理想的情形。

标准偏差的预估值以及均值和标准偏差两者的试验值，这三种情形均在图 5.3 中给出。虽然预估值和试验值之间的数值吻合并不总是很明显，但是定性吻合却处处存在。因此，看来即使在相对复杂的情况下统计理论也可以产生定性的正

确结果.

5.4-6. 增长参数的不等式. 考虑到在舍入误差的统计理论和离散误差的非统计的渐近理论中由增长参数 λ_μ 引起的主要作用，自然产生这样的问题：是否不可能以这样一种方法选择多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$，即这些参数都有某个象 $+1$ 那样无害的值. 本节我们将建立某些结果，证明如果不降低方法的阶数，这是不可能做到的.

如果 ζ 是 $\rho(\zeta)$ 的一个本性根，那么对应的增长参数由关系式

$$\lambda = \frac{\sigma(\zeta)}{\zeta\rho'(\zeta)} \tag{5-263}$$

所确定(为方便起见，我们略去上标). 从关系式

$$\rho(\zeta) = (\zeta + 1)^k r(z),$$
$$\sigma(\zeta) = (\zeta + 1)^k s(z),$$

其中 $z = (\zeta - 1)/(\zeta + 1)$，我们得到

$$\rho'(\zeta) = k(\zeta + 1)^{k-1} r(z) + 2(\zeta + 1)^{k-2} r'(z). \tag{5-264}$$

如果 $\zeta \neq -1$，可以根据 $r(z)$ 和 $s(z)$ 把 λ 表示为

$$\lambda = \frac{2s(z)}{(1 - z^2) r'(z)}. \tag{5-265}$$

如果根 $\zeta = -1$，那么 $r(z)$ 的精确次数为 $k - 1$. 在(5-264)中令 $\zeta \to -1$（或等价于 $z \to \infty$）并代入(5-263)，我们得到

$$\lambda = 2b_k/a_{k-1}, \tag{5-266}$$

其中 a_{k-1} 和 b_k 分别是多项式 $r(z)$ 和 $s(z)$ 的首项系数.

如果 $r(z)$ 的根都在虚轴上，并且 $s(z)$ 由(5-118)所确定，那么 $r'(z)$ 和 $s(z)$ 或者都为奇或者都为偶，从而(5-265)右端的表达式是实的. (5-266)也同样如此. 于是在所叙述的假

设下，我们发现这些增长参数都是实的 (Dahlquist [1959], p. 39).

如果 $r(z)$ 的根都是纯虚数，那么 $\zeta=-1$ 是 $\rho(\zeta)$ 的一个根当且仅当 k 为偶数. 如果方法是最佳的，那么我们可以从 (5-121) 中确定出 b_k，并且获得对应的增长参数的值

$$\lambda=\frac{2(c_2a_{k-1}+c_4a_{k-3}+\cdots+c_ka_1)}{a_{k-1}}. \qquad (5\text{-}267)$$

根据 (5-115) 和 (5-112)，有不等式[1]

$$\lambda\leqslant 2c_2=-\frac{1}{3}, \qquad (5\text{-}268)$$

$$\lambda\leqslant 2c_k(a_1/a_{k-1}). \qquad (5\text{-}269)$$

如果连同不等式 (5-124) 来看误差常数，那么最后的结果表明 C 和 λ 不能够同时达到它们的上界 (5-124) 和 (5-269) (见问题 37).

根据上面的结果，增长参数的负值似乎可以通过使得 k 为奇数来避免. 但正如下面的定理所表明的那样，这是行不通的，因为定理并没有对 k 的奇偶性作假设.

定理 5.16. 如果 $m=k\geqslant 2$，并且如果 $\sigma(\zeta)$ 取成使阶数达到最大，那么

$$\sum_{\mu=2}^{k}\lambda_\mu<0. \qquad (5\text{-}270)$$

证明是通过用两种不同的方式计算多项式 (5-137) 的根 ζ_μ 的乘积来完成的. 根据 Vieta 公式，我们有

$$\prod_{\mu=1}^{k}\zeta_\mu=(-1)^k\frac{\alpha_0-hA\beta_0}{\alpha_k-hA\beta_k}$$

$$=(-1)^k\frac{\alpha_0}{\alpha_k}\left\{1+hA\left(\frac{\beta_k}{\alpha_k}-\frac{\beta_0}{\alpha_0}\right)+O(h^2)\right\}. \qquad (5\text{-}271)$$

---

1) 对于 (5-268)，参阅 Dahlquist [1959], p. 40.

从 (5-141) 我们得到

$$\prod_{\mu=1}^{k} \zeta_\mu = \prod_{\mu=1}^{k} \zeta_\mu \prod_{\mu=1}^{k} [1 + \lambda_\mu Ah + O(h^2)]$$

$$= (-1)^k \frac{\alpha_0}{\alpha_k} \left\{1 + Ah \sum_{\mu=1}^{k} \lambda_\mu + O(h^2)\right\}. \quad (5\text{-}272)$$

比较 (5-271) 和 (5-272) 右端的 $O(h)$ 项，我们得到

$$\sum_{\mu=1}^{k} \lambda_\mu = \frac{\beta_k}{\alpha_k} - \frac{\beta_0}{\alpha_0}. \quad (5\text{-}273)$$

按照 $\zeta = -1$ 是否为根，或者由上面的评述按照 k 是偶数还是奇数，我们看到 $\prod_{\mu=1}^{k}\zeta_\mu$ 为 -1 或者 $+1$. 于是再次使用 Vieta 公式在每种情形都得到 $\alpha_0 = -\alpha_k$. 因此

$$\sum_{\mu=1}^{k} \lambda_\mu = \frac{\beta_k + \beta_0}{\alpha_k}. \quad (5\text{-}274)$$

从 (5-112) 我们得到

$$\beta_k = \beta_0 = b_0 + b_2 + \cdots + b_i,$$
$$\alpha_k = a_1 + a_3 + \cdots + a_j,$$

这里 $i = k$, $j = k - 1$, 当 k 为偶数时; $i = k - 1$, $j = k$, 当 k 为奇数时. 所以

$$\sum_{\mu=1}^{k} \lambda_\mu = 2\,\frac{a_1c_0 + (a_3c_0 + a_1c_2) + \cdots + (a_{i+1}c_0 + \cdots + a_1c_i)}{a_1 + a_3 + \cdots + a_j}$$

$$\leqslant 2c_0 + \frac{c_2(a_1 + \cdots + a_{i-1}) + \cdots + c_ia_1}{a_1 + a_3 + \cdots + a_j}. \quad (5\text{-}275)$$

由于 $\lambda_1 = 2c_0 = 1$, 利用 (5-115) 连同

$$c_{2\nu} < 0 (\nu = 1, 2, \cdots)$$

便得 (5-270).

5.4-7. **具有整系数的稳定算子.** 由 §5.4-4 的讨论得到，对于定点运算来说，如果常数 $\alpha_\mu/\alpha_k (\mu = 0, 1, \cdots, k-1)$ 为整数，那么局部舍入误差最小. 这个结论对部分双倍位精确

度和双倍位精确度运算以及在定性的意义上，甚至对于浮点运算都是成立的．本节我们将研究并解决确定给定的次数为 k 的全体多项式 $\rho(\zeta)$ 的问题，这种多项式满足稳定性条件，适合相容性条件并且具有上述性质．

不失一般性，可设 $\alpha_k = 1$，于是所要的多项式为

$$\rho(\zeta) = \zeta^k + \alpha_{k-1}\zeta^{k-1} + \cdots + \alpha_0, \qquad (5\text{-}276)$$

其中系数 $\alpha_{k-1}, \cdots, \alpha_0$ 为整数，首项系数为 1 且整系数的多项式称为首一多项式[1]．为简短起见，我们称满足定理 5.5 中稳定性条件的多项式为稳定多项式．于是有

定理 5.17． 令 $\rho(\zeta)$ 为首一稳定多项式，则存在一个整数 M，使得 $\rho(\zeta)$ 的所有非零根满足 $\zeta^M = 1$．

证[2] 如果 $\rho(\zeta)$ 只有零根，那么定理没有意义．如果 $\rho(\zeta)$ 有非零根，我们可设零根已经除掉．假设

$$\rho(\zeta) = \zeta^k + \alpha_{k-1}\zeta^{k-1} + \cdots + \alpha_0,$$

则有 $\alpha_0 \neq 0$． 设 $\rho(\zeta)$ 的根为 $\zeta_1, \zeta_2, \cdots, \zeta_k$． 由假设，所有 $|\zeta_\mu| \leqslant 1$；但因 $\alpha_0 = (-1)^k \Pi \zeta_\mu$ 为一非零整数，故

$$|\zeta_\mu| = 1, \quad \mu = 1, 2, \cdots, k.$$

现在考察函数

$$f(\zeta) = (-1)^k \rho(\zeta)\rho(-\zeta).$$

这是一个 ζ 的偶函数，而且也是次数为 $2k$ 的首一多项式．因此，它可以写成变量为 ζ^2 的 k 次首一多项式：$f(\zeta) = \rho^{(1)}(\zeta^2)$．由 $\rho(\zeta)$ 的乘积展开式，我们知道

$$f(\zeta) = (\zeta^2 - \zeta_1^2)(\zeta^2 - \zeta_2^2)\cdots(\zeta^2 - \zeta_k^2).$$

从而多项式 $\rho^{(1)}(\zeta)$ 的根是 $\zeta_1^2, \zeta_2^2, \cdots, \zeta_k^2$． 利用递推关系

---

1) 完整的代数描述为“全体整数域上的首多项式”；见 Birkhoff 及 Maclane [1953]，第 III 章．

2) 这个证明是 E. G. Straus 建议的．他把这个定理归于 Kronecker．该证明在减弱的形式，即 $\rho(\xi)$ 没有模超过 1 的根的情形下，只用了稳定性假设．

$$\rho^{(m+1)}(\zeta) = (-1)^k \rho^{(m)}(\sqrt{\zeta})\rho^{(m)}(-\sqrt{\zeta}),$$
$$m = 0, 1, 2, \cdots,$$
我们可以类似地构造一个具有根 $\zeta_\mu^{2^m}(\mu = 1, 2, \cdots, k)$ 的首一多项式 $\rho^{(m)}(\zeta)$ 的无穷序列，$m = 1, 2, \cdots$，但是，只存在有限多个次数固定为 k 次的不同的首一多项式，它们的根有单位模．这容易从下面的事实得到，即所有这种多项式的系数是有界的，因为是模为 1 的 k 个变量的初等对称函数[1]．因此，在多项式 $\rho^{(m)}(\zeta)$ 中有限多个不同的多项式只有有限多个不同的根． 所以 k 个序列 $\{\zeta_\mu^{2^m}\}(\mu = 1, \cdots, k)$ 中每一个都必须有重复的元素．从而，对适当的 m 和 $n > m$（也许还依赖于 μ）
$$\zeta_\mu^{2^n} = \zeta_\mu^{2^m} \text{ 或 } \zeta_\mu^{2^n - 2^m} = 1.$$
于是，对于 $M_\mu = 2^n - 2^m$，ζ_μ 满足 $\zeta_\mu^{M_\mu} = 1$． 如果我们令
$$M = \Pi M_\mu,$$
那么每个根满足 $\zeta_\mu^M = 1$． 这就完成了定理 5.17 的证明．

由定理 5.17 可知，$\rho(\zeta)$ 是多项式 $\zeta^M - 1$ 的一个因子．换言之，$\zeta^M - 1 = \rho(\zeta)\pi(\zeta)$，其中 $\pi(\zeta)$ 是某个次数为 $M-k$ 的多项式．通过作通常的长除法容易知道，它为整系数多项式． 现在我们求首一多项式[2]的唯一分解定理． 根据这个定理，每个首一多项式能够分解成不可约首一多项式的乘积（即首一多项式不能进一步分解成首一多项式），并且这种分解直到该多项式中因子的阶数都是唯一的． 特殊的首一多项式 $\zeta^M - 1$ 的分解是众所周知的，它由
$$\zeta^M - 1 = \prod_{d|M} \Phi_d(\zeta) \qquad (5\text{-}277)$$

---

1) 参阅 Birkoff 和 Maclane [1953], p. 146.

2) 参阅 Van der Waerden [1950], p. 75, 或 Birkhoff 和 Maclane [1953], p. 76.

给出．这里符号 d/M 表示下标 d 遍及 M 的因子，而 $\Phi_d(\zeta)$ 表示第 d 个割圆多项式．这种多项式称为首一多项式，其根是那些第 d 个单位根，这些根具有 $\zeta^m \neq 1$ 的性质，$0 < m < d$（即所谓第 d 个本原单位根）． 我们知道 Φ_d 的次数由 Euler 函数 $\varphi(d)$ 给出，它是用下面的方法确定的． 在素因子中 d 的分解由

$$d = p_1^{\nu_1} p_2^{\nu_2} \cdots p_m^{\nu_m}$$

给出，于是

$$\varphi(d) = p_1^{\nu_1 - 1}(p_1 - 1) p_2^{\nu_2 - 1}(p_2 - 1) \cdots p_m^{\nu_m - 1}(p_m - 1). \tag{5-278}$$

例如，$\varphi(2) = 1$，$\varphi(6) = 2$，$\varphi(13) = 12$．把分解（5-277）与前面 $\zeta^M - 1$ 的分解相对照，根据分解成首一多项式的唯一性，我们发现 $\rho(\zeta)$ 必须是

$$\rho(\zeta) = \Pi \Phi_{d_i}(\zeta), \tag{5-279}$$

其中 $\Sigma\varphi(d_i) = k$． 如果 $\rho(\zeta)$ 满足稳定性条件，那么任何一个 Φ_{d_i} 至多可以在乘积（5-279）中出现一次，因为否则在单位圆上就有重根了．如果 $\rho(\zeta)$ 适合相容性，那么一定含有因子 $\zeta - 1$，它恰好是第一个割圆多项式 $\Phi_1(\zeta)$． 因此我们证明了：

定理 5.18．任何适合相容性的首一稳定多项式可以写成 $\zeta^m \rho(\zeta)$，其中 $\rho(\zeta)$ 是包含 $\Phi_1(\zeta)$ 的不同的割圆多项式的乘积．

为了真正构造出一个给定次数为 k 的全体多项式 $\rho(\zeta)$，我们需要一个次数 $\leqslant k$ 的全体割圆多项式 $\Phi_d(\zeta)$ 的表．由

$$\varphi(k) \geqslant k^{1/2}, \text{ 对 } k > 6, \tag{5-280}$$

这个表是容易构造的．

于是，如果 $\rho(\zeta)$ 的次数是 7， 那么只能用 $d \leqslant 36$ 的割圆多项式 $\Phi_d(\zeta)$［因为 $\rho(\zeta)$ 也含有 $\zeta - 1$］．（5-280）的证明

表 5.17　割圆多项式

$$\varphi(d)=1\begin{cases}\Phi_1(\zeta)=\zeta-1,\\ \Phi_2(\zeta)=\zeta+1,\end{cases}$$

$$\varphi(d)=2\begin{cases}\Phi_3(\zeta)=\zeta^2+\zeta+1,\\ \Phi_4(\zeta)=\zeta^2+1,\\ \Phi_6(\zeta)=\zeta^2-\zeta+1,\end{cases}$$

$$\varphi(d)=4\begin{cases}\Phi_5(\zeta)=\zeta^4+\zeta^3+\zeta^2+\zeta+1,\\ \Phi_8(\zeta)=\zeta^4+1,\\ \Phi_{10}(\zeta)=\zeta^4-\zeta^3+\zeta^2-\zeta+1,\\ \Phi_{12}(\zeta)=\zeta^4-\zeta^2+1.\end{cases}$$

表 5.18　偶次（≤6）的稳定，首一且相容的多项式 $\rho(\zeta)$

$$k=2\quad \rho(\zeta)=\zeta^2-1$$

$$k=4\begin{cases}\rho(\zeta)=\zeta^4+\zeta^3-\zeta-1\\ \rho(\zeta)=\zeta^4-1\\ \rho(\zeta)=\zeta^4-\zeta^3+\zeta-1\end{cases}$$

$$k=6\begin{cases}\rho(\zeta)=\zeta^6+\zeta^5-\zeta-1\\ \rho(\zeta)=\zeta^6-\zeta^4+\zeta^2-1\\ \rho(\zeta)=\zeta^6-\zeta^5+\zeta-1\\ \rho(\zeta)=\zeta^6-2\zeta^4+2\zeta^2-1\\ \rho(\zeta)=\zeta^6+\zeta^5+\zeta^4-\zeta^2-\zeta-1\\ \rho(\zeta)=\zeta^6-1\\ \rho(\zeta)=\zeta^6-\zeta^5+\zeta^4-\zeta^2+\zeta-1\end{cases}$$

容易由(5-278)得到．对任意的 d 我们可以给出 Φ_d 的显式公式(参阅 Van der Waerden [1950], p. 120)．对于我们的目的来说，制出 $\varphi(d)\leqslant 4$ 的全部 $\Phi_d(\zeta)$ 的表就够了．

由表 5.17 容易得到次数 $\leqslant 6$ 的全体 $\rho(\zeta)$ 的表 [$d>2$，$\varphi(d)$ 为偶，从而没有 d 使得 $\varphi(d)=5$]．在表 5.18 中列出了这些偶次多项式．

5.5. 问题及附注

§5.1

1. (a) 利用对任意值 $z_p, z_{p-1}, \cdots, z_{p-q}$ 适用的插值多项式的 Newton 表达式，导出

$$\sum_{i=m}^{q}(-1)^{m+i}\binom{q}{i}\binom{i}{m}=\begin{cases}1, & m=q,\\ 0, & m<q.\end{cases}$$

(b) 用 $\varphi(h)\cos x$ 和 $\psi(h)\cos x$ 的形式分别表示

$$\nabla^2\cos(x+h) \text{ 和 } \nabla^4\cos(x+2h),$$

并证明

$$\lim h^{-2}\varphi(h)=-1, \quad \lim h^{-4}\psi(h)=1.$$

2. 用三次内插多项式导出 Simpson 的$\frac{3}{8}$公式

$$y_p-y_{p-3}=\frac{3}{8}h(f_p+3f_{p-1}+3f_{p-2}+f_{p-3}),$$

并证明其误差与 Simpson $\frac{1}{3}$ 公式同阶.

3. 利用生成函数，证明

$$\kappa_m=2\gamma_m-\gamma_{m-1}, \quad \kappa_m^*=2\gamma_m^*-\gamma_{m-1}^*,$$
$$m=0,1,2,\cdots.$$

4. 用步长 $h=0.1$，由

(a) 以 Runge-Kutta 方法确定初值 y_{-1} 和 y_{+1}，

(b) 取 $q=3$，用适当的 Adams-Bashforth 公式作预估，用 Adams-Moulton 公式继续计算。求初值问题

$$y'=1-xy^2, \quad y(0)=0$$

的数值解.

5. 解初值问题 $y'=x-y^2$，$y(0)=0$，用步长 $h=0.1$，

由幂级数法确定初始值 $y(-0.1)$ 和 $y(0.1)$. 由 Milne-Simpson 公式继续计算,预估公式为

(a) Adams-Bashforth 公式,取 $q=2$;

(b) Milne 预估公式 (5-34).

[答: $y_{-1}=0.0050005004$, $y_1=0.0049994997$,

$y_{10}=0.4555447054$]

6. 设 $P_{2q}(x)$ 为次数 $\leqslant 2q$ 的内插多项式,其内插点为 x_{-q}, x_{-q+1}, $\cdots$, x_q. 证明 $P_{2q}(x)$ 的偶部可表成

$$\frac{1}{2}(P_{2q}(x)+P_{2q}(-x))=\sum_{m=0}^{q}\binom{s+m}{2m}\nabla^{2m}f_m,\quad s=\frac{x}{h}.$$

由此推出下面 Simpson 公式的推广[1]

$$y_1-y_{-1}=h\sum_{m=0}^{q}\sigma_m\nabla^{2m}y'_m+R,\qquad (5\text{-}281)$$

其中

$$\sigma_m=\int_{-1}^{1}\binom{s+m}{2m}ds,\quad m=0,1,2,\cdots,$$

$$R=\sigma_{q+1}y^{(2q+3)}(\xi)h^{2q+3},\quad x_{-q}<\xi<x_q.$$

[应用在推导 (5-42) 中使用过的方法]

7*. 证明在问题 6 中所确定的系数 σ_m 满足

$$\sum_{m=0}^{\infty}\sigma_m t^{2m}=\frac{t\sqrt{1+\frac{1}{4}t^2}}{\log\left(\sqrt{1+\frac{1}{4}t^2}+\frac{1}{2}t\right)},\qquad (5\text{-}282)$$

并导出递推关系 $\sigma_0=2$,

$$\sigma_m=2^{1-2m}\binom{\frac{1}{2}}{m}-\frac{1}{3}\binom{-\frac{1}{2}}{1}2^{-2}\sigma_{m-1}-\frac{1}{5}\binom{-\frac{1}{2}}{2}2^{-4}\sigma_{m-2}$$

---

1) $q>1$ 不适用于初值问题的解,为什么?

$$-\cdots-\frac{1}{2m+1}\binom{-\frac{1}{2}}{m}2^{-2m}\sigma_0,\quad m=1,2,\cdots.$$

[利用表达式

$$\sum_{m=0}^{\infty}\binom{s+m}{2m}t^{2m}=\frac{1}{2\sqrt{1+\frac{1}{4}t^2}}\left\{\left(\sqrt{1+\frac{1}{4}t^2}+\frac{1}{2}t\right)^{2s+1}+\left(\sqrt{1+\frac{1}{4}t^2}-\frac{1}{2}t\right)^{2s+1}\right\}$$

和关于 s 的逐项积分]

§5.2

8. 求下列差分方程的通解 ($p=1,2,\cdots$):

(a) $y_{n+2}=y_{n+1}+y_n$; (b) $y_n-y_{n-p}=0$;

(c) $\nabla^p y_n=0$, 并求满足条件 $y_0=0$, $y_1=1$ 的 (a) 的解. [这样生成的序列被称为 Fibonacci 序列]

9. 求差分方程

$$\alpha_k y_{n+k}+\alpha_{k-1}y_{n+k-1}+\cdots+\alpha_0 y_n=\zeta^n$$

的一个特解,其中 ζ 是特征多项式

$$\rho(\zeta)=\alpha_k\zeta^k+\alpha_{k-1}\zeta^{k-1}+\cdots+\alpha_0$$

的一个 $p(p\geqslant 0)$ 重根.

10. 假定 $\alpha_{0,n}\neq 0,\alpha_{k,n}\neq 0$,证明线性差分方程

$$\alpha_{k,n}y_{n+k}+\alpha_{k-1,n}y_{n+k-1}+\cdots+\alpha_0 y_n=0$$

在 $n=n_0$ 处独立的 m 个解系在某个 $n>n_0$ 不能突然变成相关.

11. 设 $f(x)$ 在闭区间 $[x_{n-q}, x_n]$ 内为 q 次连续可微函数,证明对该区间内的某个 ξ, 有

$$f^{(q)}(\xi)=h^{-q}\nabla^q f_n.$$

[q 次应用 Rolle 定理于函数 $f(x)-P(x)$, 其中 $P(x)$ 由 (5-

12) 给定]作为一个应用，证明

$$\lim h^{-k}\nabla^k\zeta^n = n(n-1)\cdots(n-k+1)\zeta^{n-k}. \tag{5-283}$$

12. 对于 $n = 0$ 的序列 (5-70)，求矩阵 (5-62) 的行列式的值. 方法如下：

(a) 使用由

$$y_n = h^{-k}\nabla^k\zeta_\mu^n \quad (\mu = 1,2,\cdots,m;\ k = 0,1,\cdots,p_\mu - 1)$$

所确定的组代替 (5-70)，这里 $h \neq 0$；

(b) 通过行的初等变换，将该行列式化成以

$$h^{-k}(\zeta_\mu - kh) \quad (\mu = 1,2,\cdots,m; k = 0,1,\cdots,p_\mu - 1)$$

为元素的 Vandermonde 行列式；

(c) 利用 (5-283)，令 $h \to 0$.

13*. 令 $a_\mu, \varphi_\mu, \delta_\mu (\mu = 1, 2,\cdots,k)$ 为实数，$0 < \varphi_\mu \leqslant \pi$，令所有数 φ_μ 互异，并设至少有一个 a_μ 异于零. 证明序列 $\{s_n\}$ 当 $n \to \infty$ 时无极限，这里

$$s_n = a_1\cos(n\varphi_1 + \delta_1) + \cdots + a_k\cos(n\varphi_k + \delta_k).$$

14. 用迭代法解关于 y 的 Kepler 方程

$$x = y - \varepsilon\sin y,$$

其中 $x = 0.8$，$\varepsilon = 0.2$，初值取为 $y^{(0)} = x$. 在第五步后应用定理 5.4 来估计误差. 与解析解

$$y = x + \sum_{n=1}^{\infty} \frac{2}{n} J_n(n\varepsilon)\sin nx \quad (J_n = n \text{ 阶 Bessel 函数})$$

进行比较.

15. 证明与差分方程

$$y_{n+4} - y_n = \frac{4}{3}h(2f_{n+3} - f_{n+2} + 2f_{n+1}), \tag{5-284}$$

$$y_{n+3} - y_n = \frac{3}{8}h(f_{n+3} + 3f_{n+2} + 3f_{n+1} + f_n) \tag{5-285}$$

有关的算子都是4阶的，并证明它们的误差常数分别为 $\frac{7}{90}$ 和 $-\frac{1}{80}$ {分别在点 x_{n+2} 和 $x_{n+3/2}$ 附近展开 $L[y(x);h]$}.

16. 用使得与

$$y_{n+3}-y_n+\alpha(y_{n+2}-y_{n+1})=h\beta(f_{n+2}+f_{n+1})$$

有关的算子为4阶的方法来确定常数 α,β，并确定误差常数. 验证由此得到的算子是不稳定的.

17. 通过直接数值计算证明按上述问题得到的不稳定算子不能够用来解问题5中所叙述的初值问题，但是它作为与 Milne 公式连用的预估要比问题5中提出的预估公式为好[1].

18. 确定常数 α,β 和 γ，使得与显式差分方程

$$y_{n+2}-y_{n-2}+\alpha(y_{n+1}-y_{n-1})=h[\beta(f_{n+1}-f_{n-1})+\gamma f_n]$$

有关的算子为六阶. 所得的算子稳定吗？（答 $\alpha=28$，$\beta=12$，$\gamma=36$；否！）

19. 关于稳定性的充分条件. 设 $\rho(1)=0$，并设多项式

$$\rho_1(\zeta)=\frac{\rho(\zeta)}{\zeta-1}=\gamma_{k-1}\zeta^{k-1}+\gamma_{k-2}\zeta^{k-2}+\cdots+\gamma_0$$

的系数满足 $\gamma_{k-1}>\gamma_{k-2}>\cdots>\gamma_0$. 证明 $\rho(\zeta)$ 满足稳定性条件. 应用这个结果证明基于数值微分的方法(5-48)对于 $r=0$ 和 $m=1,2,3,4$ 是稳定的.

20*. 建立在数值微分基础上的方法(5-48)是 q 阶的. 通过把一个适合的项加到多项式 $\sigma(\zeta)$ 上使它为 $q+1$ 阶 [答: $\sigma(\zeta)=\zeta^{q-r}-\delta_{r,q+1}(\zeta-1)^q$].

21*. 假设由(5-120)确定的系数 c_{2n} 满足 $\sum_{n=0}^{\infty}c_{2n}=0$，证明 Dahlquist 定理（参阅 Dahlquist [1956], p. 52），即对任何

---

1) 由 J. Titus 完成的数值试验表明，不稳定算子可以可靠地用作预估公式.

显式算子 $p \leqslant k$. [说明 $p > k$ 时意味着 $\beta_k = s(1) > 0$]

22.确定对应于多项式

(a) $\rho(\zeta) = \zeta^6 - 2\zeta^4 + 2\zeta^2 - 1$; (b) $\rho(\zeta) = \zeta^6 - \zeta^4 + \zeta^2 - 1$ 的最佳算子。

23.证明由 (5-134) 确定的系数 k_{2n} 满足递推关系

$$k_{2n} + \frac{1}{3}\binom{-\frac{1}{2}}{1}2^{-2}k_{2n-2} + \cdots + \frac{1}{2n+1}\binom{-\frac{1}{2}}{n}2^{-2n}k_0 = \binom{\frac{1}{2}}{n}2^{-2n-1}. \quad n = 0,1,2,\cdots.$$

§5 3

24.对于 Simpson 的 $\frac{3}{8}$ 公式 (5-285) 以及 Milne 预估公式 (5-34) 计算增长参数 λ_μ.

25*. 证明：在用 Milne 方法解 $y' = Ay$, $y(0) = 1$ 中的不稳定分量 $(-1)^n e^{-Ax_n/3}$ 可以通过下面的方法来消除，即，对每一个固定的次数（例如，对于 $n = mp$, p 为整数，$m = 1, 2, \cdots$）把由 Milne 方法获得的值与由 Simpson 的 $\frac{3}{8}$ 公式 (5-285) 所得到的值作平均（参阅 Milne 和 Reynolds [1959]）。

26.对于多项式

$$\text{(a) } \rho(\zeta) = -\frac{3}{2}\zeta^2 + 2\zeta \quad \frac{1}{2};$$

$$\text{(b) } \rho(\zeta) = (\zeta^2 \quad 1)(\zeta^2 - 2\zeta\cos\alpha + 1)$$

确定在引理 5.5 中所定义的常数 Γ. 证明对于

$$\rho(\zeta) = (\zeta - 1)^2, \text{ 有 } \Gamma = \infty.$$

27.证明在差分算子 $L[y(x); h]$ 的积分表示式 $(5\text{-}17^8)$ 中的核 $G(s)$ 满足

$$\int_0^k G(s)ds = C_{p+1},$$

其中 C_{p+1} 是 §5.2-5 中所定义的常数.

28. 如果核 $G(s)$ 在区间 $[0,k]$ 上不改变符号，则称算子 $L[y(x);h]$ 为有定的. 证明广义的中值定理 (5-177) 对一切具有连续的 $y^{(p+1)}(x)$ 的函数 $y(x)$ 均成立的充分必要条件是算子 $L[y(x);h]$ 为有定的.

29. 证明问题 15 中所考虑的算子是有定的.

30. 关于有定的必要条件. 如果算子 $L[y(x);h]$ 是有定的，又如果 $\beta_k = 0$，证明 $\operatorname{sign} \alpha_k = \operatorname{sign} C_{p+1}$. 如果该算子是有定的并且 $\beta_k \neq 0$，证明 $p\beta_k/\alpha_k$ 位于区间 $[0,1]$ 的外部，而且

$$\operatorname{sign} \beta_k = -\operatorname{sign} C_{p+1}.$$

31. 对下面的特殊情形作出定理 5.11 的估计: (a) 取 $q = 2$ 的 Milne-Simpson 方法; (b) 取 $q = 3$ 的 Adams-Bashforth 方法; (c) 取 $q = 4$ 的 Adams-Moulton 方法.

32. 改进的外推到极限. 在这个问题和下面的二个问题中，我们考察偶阶 $k = 2s$ 的对称(但未必最佳)差分方程. 设

$$\begin{aligned} &\alpha_s y_{n+s} + \alpha_{s-1} y_{n+s-1} + \cdots + \alpha_{-s} y_{n-s} \\ &\quad = h\{\beta_k f_{n+s} + \beta_{s-1} f_{n+s-1} + \cdots + \beta_{-s} f_{n-s}\}, \end{aligned} \tag{5-286}$$

其中 $\alpha_\mu = -\alpha_{-\mu}$, $\beta_\mu = -\beta_{-\mu}$, $\mu = 0, 1, \cdots, s$, 便是这样的差分方程. 证明与 (5-286) 有关的差分算子 $L[y(x);h]$ 的阶数 p 为偶数，并且

$$L[y(x);h] = C_{p+1} h^{p+1} y^{(p+1)}(x) + O(h^{p+3}).$$

33*. 设由形如 (5-286) 的稳定和相容的差分算子解初值问题 $y' = f(x,y)$, $y(a) = \eta$, 并设给定的开始值为 $y_{-s}, y_{-s+1}, y_{s-1}$, 于是取 $n = 0$ 首次应用 (5-286). 如果开始误差 e_μ 是 $O(h^{p+1})$, $\mu = -s, -s+1, \cdots, s-1$, 那么由定理 5.12 得到

$$e_n = h^p e(x_n) + O(h^{p+1}), \tag{5-287}$$

其中

$$e(x) = g(x)e(x) - Cy^{(p+1)}(x), e(a) = 0, \tag{5-288}$$

$$C = \frac{C_{p+1}}{\beta_{-s} + \beta_{-s+1} + \cdots + \beta_s}.$$

证明：如果开始误差满足

$$e_\mu = h^p e(x_\mu) + O(h^{p+2}), \tag{5-289}$$

其中 $e(x)$ 由 (5-288) 确定，那么关系式 (5-287) 可以改进到

$$e_n = h^p e(x_n) + O(h^{p+2}) \tag{5-290}$$

[提示：利用问题 32 的结果，导出一个关于量

$$z_n = e_n - h^p e(x_n)$$

的递推关系，而对这个量可以应用引理 5.6，取 $Z = O(h^{p+2})$ 和 $A = O(h^{p+1})$] 结果 (5-290) 意味着应用 Richardson 延迟趋向于极限对适当的开始解，将提高近似解的阶二个单位[1] (实际上，这个结论只有当舍入误差可以忽略时才成立)。

34. 证明条件 (5-289) 对这些开始值

$$y_\mu = y(x_\mu) - Ch^{p+1}\mu y^{(p+1)}(a), \quad \mu = -s, -s+1, \cdots s-1 \tag{5-291}$$

成立。

35*. 如果取中点公式为预估公式来使用梯形公式，并且只作一次校式，由此得到公式

$$y_{n+1} - y_n = \frac{1}{2}h\{f_n + f^*_{n+1}\}, \tag{5-292}$$

其中 $f_n = f(x_n, y_n)$，$f^*_{n+1} = f(x_{n+1}, y_{n-1} + 2hf_n)$。 根据 §5.3-7 的结果，这个方法的渐近误差和用无限次校正的梯形公式相同。 如果无损于精确度，对于 $n \geqslant 1$，(5-292) 可用公式

1) 在这方面，对于特殊而不够完善的结果，参阅 Gaunt[1927] 和 de Vogelaere [1957]。

$$y_{n+1} - y_n = \frac{1}{2} h\{f_n^* + f_{n+1}^*\} \tag{5-293}$$

来代替，该公式每步只要求计算一次 $f(x, y)$ 的值，这样对吗?［否；考察例子 $y' = Ay$, $y(0) = 1$］

§5.4

36. 对于问题

$$y' = -\frac{1}{x} y, \quad y(1) = 1,$$

当

(a) $q = 1$（定点运算）;

(b) $q = y^2$（浮点运算）

时，用 Milne-Simpson 求积确定函数 $v_\mu(x)$。

37*. 已经证明了属于 $\zeta = -1$ 的最大算子的增长参数 λ 和误差常数 C 满足不等式

$$\lambda \leqslant -\frac{1}{3}, \ (5\text{-}268); \quad C \leqslant 2^{-k-1} c_{k+2}. \tag{5-125}$$

如果 $A = \lambda + \frac{1}{3}$, $B = C - 2^{-k-1} c_{k+2}$, 证明对于 $k \geqslant 4$

$$AB \geqslant 2^{-k-1} c_4 c_k. \tag{5-294}$$

［该结果表明 λ 和 C 的界虽然分别是严格的，却不能同时逼近］

38. 虽然导出 (5-268) 的这些假设对于 $q \geqslant 3$ 是不满足的，证明该不等式对于推广的 Milne-Simpson 方法仍然成立。（多项式 $r(z)$ 的系数仍然是正的。）

39. 对于一个阶 $p = k \geqslant 2$ 的显式方法来说，$\zeta = -1$ 为 $\rho(\zeta)$ 的一个根，且对应的增长参数 $\leqslant -1$。［Dahlquist, oral communication.］

40. 确定全体符合相容性的五阶首一稳定多项式。

注

§5.1-2 和 §5.1-3. 其它的线性多步公式已经由 Capra [1956], Keitel [1956], Löwdin [1952], Sconzo [1954] 提出. 变步长的线性多步方法参阅 Urabe 和 Mise [1955]. 有关方法,严格地说不是这里所考虑的线性多步类型, 由 Urabe 和 Tsushima [1953], Wilf [1957], Wall [1956] 给出. Adams 方法的计算方面由 Collatz [1960], p. 78 以及 Ralston [1960] 予以讨论. 在第五章讨论过的所有方法, 实质上是根据这种思想,即,解用多项式逼近是最好的. 对于某些微分方程, 特别是那些具有高度振动分量的解,这种逼近可能是不适合的. 而用某些其它简单的函数类,象指数函数,可能较好. 这种逼近由 Brock 和 Murray [1952] 进行讨论. 也可参阅 Certaine [1960]. 这种思想值得进一步研究.

§5.2-1. 关于差分方程的一个更加完整的分析由 Goldberg [1958], Fort [1948], Milne-Thomson [1933], Nörlund [1924] 给出.

§5.2-3. 这里给出的收敛性的定义与 Dahlquist [1956] 所介绍的"稳定的收敛性"的概念是不同的.

§5.2-4. 这里所介绍的稳定性条件由 Todd [1950] 试探性地讨论了. 这里所用的确切定义取自 Dahlquist [1956]. 对于常微分方程数值积分不稳定的方法由 Muhin [1952a], Salzer [1956], Quade [1957], Kopal [1958] 提出. 稳定性的许多不同的定义在文献中已经提供,参看 Carr [1957], Liniger [1957], Hamming [1959], Wilf [1959], Budak 和 Gorbunov [1959], Ralston [1960], Hull 和 Luxemburg [1960]. 在这些定义中, 有些与我们的收敛性定义密切相关, 其它的与我们在 §5.3-1 中所介绍的弱或者条件稳定的概念有关. 这些定义常常只严格地用于积分 $y' = Ay$, 其中 A 是一个常数. $r = 0$ 以及 $q = 1,2,3,4,5,6$ 时的微分法的稳定性由 Mitchell 和 Craggs

[1953] 证明.

§5.2-7. 对于其它的构造差分算子的方法,见 Frei [1954], Hamming [1959], Hull 和 Newbery [1959], Robertson [1960].

§5.2-8. 一个有关的结果,参阅 Ceschino 和 Kuntzmann [1958].

§5.3-1. 一个类似而又更加特殊的计算,由 Loud [1949] 完成. 弱稳定现象由 Dahlquist [1951] 以特殊的情形报导,而关于常系数方程的综合性讨论由 Rutishauser [1952] 给出.

§5.3-3. 稳定性和相容性一起是收敛性的充分必要条件这一事实是——对线性微分方程——由 Lax 和 Richtmyer [1956] 给出更为一般的证明(也见 Richtmyer [1957]).

§5.3-4. 关于 Adams 方法的误差界有大量文献, 见 von-Mises [1930], Tollmien [1938, 1953], Fricke [1949], Hamel [1949], Weissinger [1950, 1952], Matthieu [1951], Mohr [1951], Vietoris[1953a, 1955b], Bahvalov[1955a, 1955b], Gaier[1956], Zondek 和 Sheldon[1959]. Milne 方法的误差由 Richter[1951] 估计, 而一般情形求积方法的误差由 Shura-Bura [1952] 和 Hildebrand [1956, p. 219] 作出估计. Eltermann [1955], Uhlmann [1957a], Richter [1952] 和 Dahlquist [1959] 给出了用各种线性多步方法积分一阶方程组的误差界. 也见 Hull 和 Newbery [1959].

§5.3-5. 离散误差的渐近公式由 Brodskii [1953], Lotkin [1954], Gray [1955] 给出.

§5.3-6. 用比较预估和校正值得到局部离散误差的估计值,由 Lotkin [1957] 作了非正式的讨论.

§5.3-7. 仅预估一次的算法由 Wall [1956] 提出,也可见 Young [1957] 关于这篇论文的评论.

§5.4-1. 累积舍入误差由 Bahvalov [1955a, 1955b], Da-

hlquist [1956, 1959], Lotkin [1954] 给出.

§5.4-3. 在多步方法中，舍入误差传播的统计方法由 Sterne [1953] 讨论。本节的结果在 Henrici [1960] 中发表。

第六章　二阶特殊方程的线性多步方法

第五章中所讨论的方法与所得的结果虽然只是对单个微分方程阐述的，但是正如第二章中的结果推广到第三章的方程组一样，它也可以推广到微分方程组．因此，例如要积分一个二阶微分方程

$$y'' = f(x, y, y'), \tag{6-1}$$

便可以把它化成方程组 $y' = z$, $z' = f(x, y, z)$，并且应用于第五章中所描述的方法之一．如同单步法一样，这个过程是完全合理的（既不损失精确度，也没有花费不必要的计算）．

如果要积分的方程形如

$$y'' = f(x, y), \tag{6-2}$$

即若在这个微分方程的右端不出现导数，那情形就略有不同．这种类型的方程称为特殊的微分方程，同样，形如

$$y^{(n)} = f(x, y)$$

的方程也叫做特殊微分方程，其中 n 为大于 2 的整数．二阶的特殊微分方程，特别是这样的方程组，例如，在没有耗散的力学问题中是经常出现的．如果对一阶导数没有兴趣，那么人为地把它们化为一阶方程组是不必要的．而事实上，一个多世纪以来，天文学家在积分 (6-2) 时所用的多步型方法是不使用一阶导数进行计算的．本章专门研究这些方法，虽然该理论大体上与一阶方程的多步法理论相仿，但有足够多的新的与意料不到的在误差传播领域中值得注意的性质，因此将它们的讨论另立一章是合理的．下面的讨论着重于这些新的方面，而将与一阶情形多少有些类似的方面一笔带过．对于

任意阶的特殊方程,可以建立类似的理论. 然而,由于在实际应用方面,高于二阶的特殊方程出现相对来说是少的,所以,我们只限于讨论二阶的情形.

6.1. 线性多步方法的局部研究

6.1-1. *某些特殊方法*. 我们要做的第一件事是导出一个满足

$$y''(x) = f(x, y(x)) \tag{6-3}$$

的函数的类似于(5-13)的公式. 通过两次积分,我们得到公式

$$y(x+k) - y(x) = ky'(x) + \int_x^{x+k} (x+k-t) \times f(t, y(t))dt.$$

它也可以看作是带有余项的 Taylor 公式. 对我们的目的来说,这个结果还不够理想,因为我们不希望用到一阶导数 $y'(x)$. 然而,在同一结果中把 k 换成 $-k$ 并相加就可以消去一阶导数. 记 $f(t) = f(t, y(t))$, 则二个积分的和便可以化成如下的形式:

$$\int_x^{x+k} (x+k-t)f(t)dt + \int_x^{x-k} (x-k-t)f(t)dt = \int_x^{x+k} (x+k-t)[f(t) + f(2x-t)]dt,$$

从而便得到恒等式

$$y(x+k) - 2y(x) + y(x-k) = \int_x^{k+x} (x+k-t)[f(t) + f(2x-t)]dt. \tag{6-4}$$

此恒等式是许多积分公式的基础. 当 $f(t)$ 换成在点 $x_p, \cdots, x_{p-q}$ 上的 q 次插值多项式时,按我们对 x, k 及 q 的选择,便能

表 6.1

方　　法	x	$x+h$	q
Störmer	x_p	x_{p+1}	$\geqslant 0$
Cowell	x_{p-1}	x_p	$\geqslant 2$
(6-15)	x_{p-1}	x_{p+1}	$\geqslant 2$
(6-19)	x_{p-2}	x_p	$\geqslant 4$

得到一系列的特殊方法，其中一些方法列于表 6.1 中．下面对这些方法进行仔细的讨论．

(i) Störmer 方法（Störmer [1907, 1921]）．上面梗概的说明给出

$$y_{p+1} - 2y_p + y_{p-1} = h^2 \sum_{m=0}^{q} \sigma_m \Delta^m f_p, \qquad (6-5)$$

其中

$$\begin{aligned}\sigma_m &= \frac{(-1)^m}{h^2} \int_{x_p}^{x_{p+1}} (x_{p+1} - x) \\ &\quad \times \left[\binom{-s}{m} + \binom{s}{m}\right] dx \left(s = \frac{x - x_p}{h}\right) \\ &= (-1)^m \int_0^1 (1 - s) \left[\binom{-s}{m} + \binom{s}{m}\right] ds. \qquad (6-6)\end{aligned}$$

交换求和与积分的顺序，正如 §5.1-2 中所述，我们得到

$$S(t) = \sum_{m=0}^{\infty} \sigma_m t^m = \left[\frac{t}{\log(1-t)}\right]^2 \frac{1}{1-t}. \qquad (6-7)$$

从

$$\begin{aligned}\frac{d}{dt}[\log(1-t)]^2 &= -\frac{2}{1-t}\log(1-t) \\ &= 2(1 + t + t^2 + \cdots)\left(t + \frac{1}{2}t^2 + \frac{1}{3}t^3 + \cdots\right) \\ &= 2(h_1 t + h_2 t^2 + h_3 t^3 + \cdots),\end{aligned}$$

其中 $h_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m}$ 表示调和级数的第 m 次部分和，便得到

$$\left[\frac{\log(1-t)}{t}\right]^2 = 1 + \frac{2}{3}h_2 t + \frac{2}{4}h_3 t^2 + \cdots \quad (6\text{-}8)$$

（在 §6.1-7 的不同的上下文中需要类似的展式）。用 (6-7) 乘 (6-8) 的两侧，我们得到

$$\left(1 + \frac{2}{3}h_2 t + \frac{2}{4}h_3 t^2 + \cdots\right)(\sigma_0 + \sigma_1 t + \sigma_2 t^2 + \cdots)$$
$$= 1 + t + t^2 + \cdots.$$

由此得到递推关系式

$$\sigma_0 = 1,$$
$$\sigma_m = 1 - \frac{2}{3}h_2\sigma_{m-1} - \frac{2}{4}h_3\sigma_{m-2} - \cdots - \frac{2}{m+2}h_{m+1}\sigma_0,$$
$$m = 1, 2, \cdots.$$

从这个递推式，容易得到表 6.2 中的数值.

表 6.2　σ_m 的值

m	0	1	2	3	4	5	6
σ_m	1	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{19}{240}$	$\frac{3}{40}$	$\frac{863}{12096}$

公式 (6-5) 的使用大致与 Adams-Bashfovth 公式相同. 一旦值 $y_p, \cdots y_{p-q}$ 为已知，y_{p+1} 就能显式地计算而无需迭代. 在确定开始值时，还有困难，这可用幂级数法或第四章所描述的方法之一而得到它们. 为了节省工作量以及增加精确度，建议开始点取在对 x_0 为对称的位置上. 例如，假若需要 5 个开始值，它们应定在 x_{-2}, x_{-1}, x_0, x_1, x_2 处. 当 $q = 0$ 及 $q = 1$ 时，Störmer 公式便化成简单的公式：

$$y_{p+1} - 2y_p + y_{p-1} = h^2 f_p. \quad (6\text{-}9)$$

(ii) Cowell 方法（Cowell 及 Crommelin [1910]）. 按照表 6.2，令

$$y_p - 2y_{p-1} + y_{p-2} = h^2 \sum_{m=0}^{q} \sigma_m^* \nabla^m f_p, \qquad (6\text{-}10)$$

其中

$$\sigma_m^* = \frac{(-1)^m}{h^2} \int_{x_{p-1}}^{x_p} (x_p - x) \left[\binom{-s}{m} + \binom{s+2}{m} \right] dx$$

$$= (-1)^m \int_{-1}^{0} (-s) \left[\binom{-s}{m} + \binom{s+2}{m} \right] ds. \qquad (6\text{-}11)$$

像上面一样确定系数 σ_m^* 的生成函数，其结果为

$$S^*(t) = \sum_{m=0}^{\infty} \sigma_m^* t^m = \left[\frac{t}{\log(1-t)} \right]^2. \qquad (6\text{-}12)$$

由关系式 $S^*(t) = (1-t)S(t)$ 可导出

$$\sigma_m^* = \sigma_m - \sigma_{m-1}, \quad m = 1, 2, \cdots. \qquad (6\text{-}13)$$

用展开式 (6-8) 乘以 (6-12) 的两端，便得到

$$\left(1 + \frac{2}{3} h_2 t + \frac{2}{4} h_3 t^2 + \cdots\right)(\sigma_0^* + \sigma_1^* t + \sigma_2^* t^2 + \cdots)$$

$$= 1.$$

通过比较系数，我们得到递推关系式 $\sigma_0^* = 1$,

$$\sigma_m^* = -\frac{2}{3} h_2 \sigma_{m-1}^* - \frac{2}{4} h_3 \sigma_{m-2}^* - \cdots - \frac{2}{m+2} h_{m+1} \sigma_0^*,$$

$$m = 1, 2, \cdots.$$

易得列于表 6.3 中的数值.

表 6.3 σ_m^* 的值

m	0	1	2	3	4	5	6
σ_m^*	1	-1	$\frac{1}{12}$	0	$-\frac{1}{240}$	$-\frac{1}{240}$	$-\frac{221}{60480}$

Cowell 公式不用于 $q \geqslant 0$ 的情形. 当 $q=1$ 时它化成(6-9). 当 $q=2$ 和 $q=3$ 时,我们得到通用的公式

$$y_p - 2y_{p-1} + y_{p-2} = h^2\left\{f_{p-1} + \frac{1}{12}\nabla^2 f_p\right\}$$

$$= \frac{1}{12}h^2\{f_p + 10f_{p-1} + f_{p-2}\}. \qquad (6\text{-}14)$$

$q \geqslant 2$ 时 Cowell 公式为隐式，即未知的 y_p 值不仅在左端出现，同时也是右端 f_p 中的变量. 除微分方程为线性的情形外,所产生的关于 y_p 的方程是非线性的，它可通过迭代法来求解，正如在 §5.1-2 (ii) 中所描述的. 定理 5.4 保证了迭代过程的收敛性，那里给出的误差估计式也能应用（将 h 换为 h^2). 由于 Störmer 公式的左端有相同的 y_n 的组合值，故被推荐为预估公式.

(iii) 另外的一些特殊方法. 另一些有用的公式可通过将步长 h 加倍而导出，类似于关于一阶方程的 Nyström 及 Milne-Simpson 方法的推导. 取 $k=2h$ 及 $x=x_{p-1}$，便得到公式

$$y_{p+1} - 2y_{p-1} + y_{p-3} = h^2\sum_{m=0}^{q}\tau_m\nabla^m f_p, \qquad (6\text{-}15)$$

其中系数 τ_m 由

$$\tau_m = (-1)^m\int_{-1}^{1}(1-s)\left[\binom{-s}{m} + \binom{s+2}{m}\right]ds \qquad (6\text{-}16)$$

给出,其生成函数为

$$T(t) = \sum_{m=0}^{\infty}\tau_m t^m = \left[\frac{t}{\log(1-t)}\right]^2\frac{4-4t+t^2}{1-t}. \qquad (6\text{-}17)$$

由它导出递推关系为 $\tau_0=4$, $\tau_1=-4$,

$$\tau_m = 1 - \frac{2}{3} h_2\tau_{m-1} - \frac{2}{4} h_3\tau_{m-2} - \cdots - \frac{2}{m+2} h_{m+1}\tau_0,$$

$$m = 2, 3, \cdots.$$

表 6.4 给出数值结果.

表 6.4 τ_m 的 值

m	0	1	2	3	4	5	6
τ_m	4	-4	$\frac{4}{3}$	0	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{61}{945}$

公式 (6-15) 为显式. 当 $q < 2$ 时, 该公式不被推荐. $q = 2$ 时得到简单而又相对精确的公式

$$y_{p+1} - 2y_{p-1} + y_{p-3} = \frac{4}{3} h^2\{f_p + f_{p-1} + f_{p-2}\}. \tag{6-18}$$

由于 $\tau_3 = 0$, 故选取值 $q = 3$ 也得到同样公式.

在 (6-4) 中选取 $k = 2h$, $x = x_{p-2}$, 便导出另一组公式. 我们有

$$y_p - 2y_{p-2} + y_{p-4} = h^2 \sum_{m=0}^{q} \tau_m^* \nabla^m f_p, \tag{6-19}$$

其中

$$\tau_m^* = (-1)^m \int_{-2}^{0} (-s) \left[\binom{-s}{m} + \binom{s+4}{m}\right] ds. \tag{6-20}$$

由生成函数

$$T^*(t) = \sum_{m=0}^{\infty} \tau_m^* t^m = \left[\frac{t}{\log(1-t)}\right]^2 (4 - 4t + t^2) \tag{6-21}$$

可用通常的方式导出递推公式 $\tau_0^* = 4$, $\tau_1^* = -8$,

$$\tau_2^* = 4 + \frac{4}{3}, \quad \tau_m^* = -\frac{2}{3} h_2\tau_{m-1}^* - \frac{2}{4} h_3\tau_{m-2}^* - \cdots$$

$$- \frac{2}{m+2} h_{m+1}\tau_0^*, \quad m = 3, 4, \cdots.$$

通过对(6-17)与(6-21)的比较，我们还可得到

$$\tau_m^* = \tau_m - \tau_{m-1},\quad m = 1,2,\cdots.$$

容易得到表 6.5 中的数值结果.

当 $q = 0,1,2$ 时，(6-19)具有不规则的形式，故不推荐它去解决实际问题. $q = 3$ 时该公式为

$$y_p - 2y_{p-2} + y_{p-4} = \frac{4}{3}h^2\{f_{p-1} + f_{p-2} + f_{p-3}\}.$$

因此，它与(6-18)等价. 当 $q = 4$ 时，由于 $\tau_5^* = 0$，亦是 $q = 5$ 的情形，(6-19)化成为

$$y_p - 2y_{p-2} + y_{p-4} = \frac{1}{15}h^2\{f_p + 16f_{p-1} + 26f_{p-2} + 16f_{p-3} + f_{p-4}\}. \tag{6-22}$$

还有一些公式可以通过对上述的某些公式线性组合而成，也可参看 §6.1-6.

表 6.5 τ_m^* 的 值

m	6	1	2	3	4	5	6
τ_m^*	4	-8	$\frac{16}{3}$	$-\frac{4}{3}$	$\frac{1}{15}$	0	$-\frac{2}{945}$

6.1-2. 关于特殊二阶方程的一般算子. 前节考虑的所有方法均为差分方程

$$\alpha_k y_{n+k} + \alpha_{k-1}y_{n+k-1} + \cdots + \alpha_0 y_n = h^2\{\beta_k f_{n+k} + \beta_{k-1}f_{n+k-1} + \cdots + \beta_0 f_n\} \tag{6-23}$$

的特殊情形，其中 $f_m = f(x_m, y_m)$. 这个表达式与(5-58)，除 h 被换成 h^2 这个重要的不同外，是完全相同的. 在本章的余下部分，我们将集中考虑一般的表达式(6-23)而不考虑 §6.1-1 中的任一特殊方法. 为了对用于积分 $y'' = f(x,y)$ 的(6-23)定义一个"好"的方法，我们仍试图发现必要与充分条

件．在研究的过程中，对本章开头探讨的具体方法的性质将可得到透彻的了解．在整个过程中，我们假设 $\alpha_k \neq 0$，$|\alpha_0| + |\beta_0| > 0$。于是差分方程 (6-23) 的阶 k 被唯一确定．我们指出，对于 §6.1-1 中所考虑的特殊方法，步数 k 至少为 2．

与差分方程 (6-23) 有关的差分算子为

$$\begin{aligned}L[y(x);h] = {}& \alpha_k y(x+kh) + \alpha_{k-1} y(x+(k-1)h) + \cdots \\ & + \alpha_0 y(x) - h^2\{\beta_k y''(x+kh) + \beta_{k-1} y''(x+(k-1)h) \\ & + \cdots + \beta_0 y''(x)\}。\end{aligned} \tag{6-24}$$

此算子可以作用于任何具有二阶导数的函数 $y(x)$，但目前只应用于具有充分高阶连续导数的函数上．这样我们便可将 (6-24) 展为 h 的幂级数并得到

$$\begin{aligned}L[y(x);h] = {}& C_0 y(x) + C_1 y'(x)h + \cdots \\ & + C_q y^{(q)}(x) h^q + \cdots,\end{aligned} \tag{6-25}$$

其中 $C_q(q = 0,1,2,\cdots)$ 与 $y(x)$ 无关．特别是，

$$C_0 = \alpha_0 + \alpha_1 + \cdots + \alpha_k,$$

$$C_1 = \alpha_1 + 2\alpha_2 + \cdots + k\alpha_k,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\begin{aligned}C_q = {}& \frac{1}{q!}\{\alpha_1 + 2^q\alpha_2 + \cdots + k^q\alpha_k\} \\ & - \frac{1}{(q-2)!}\{0^{q-2}\beta_0 + 1^{q-2}\beta_1 + \cdots + k^{q-2}\beta_k\} \\ & q = 2,3,\cdots.\end{aligned} \tag{6-26}$$

差分算子 (6-24) 的阶 p 定义为使 $C_q=0$，$q=0,1,\cdots p+1$；$C_{p+2} \neq 0$ 的唯一整数．与 §5.2-5 一样，易证 p 与 C_{p+2} 并不依赖于 t，假如将 (6-23) 换成

$$\alpha_k y_{n+t+k} + \cdots + \alpha_0 y_{n+t} = h^2\{\beta_k f_{n+t+k} + \cdots + \beta_0 f_{n+t}\}.$$

由定义

$$L[y(x);h] = h^{p+2} C_{p+2} y^{(p+2)}(x) + O(h^{p+3}). \tag{6-27}$$

因此，一个算子的阶可以看作是：当以 y_m 来代替 $y''=f(x,y)$ 的精确解时，(6-23) 所满足的精确度的一个量度. 在给定阶的所有差分算子类中，常数

$$C=\frac{C_{p+2}}{\beta_0+\beta_1+\cdots+\beta_k} \tag{6-28}$$

称为误差常数，它表示对精确度的更进一步的度量. 值得注意的是：对平凡算子 $L[y(x);h]$ 乘以一个非零常数时，p 与 C 为不变量.

一个算子的阶与误差常数通常无需一一计算常数 C_0, C_1, $\cdots$ 便可求得. 如果已知当 $y(x)$ 的次数不超过 $p+1$ 次多项式时 $L[y(x);h]=0$，而对一些次数为 $p+2$ 的多项式

$$L[y(x);h]\neq 0,$$

显见其阶为 p. 现在我们来看看 §6.1-1 中的特殊公式. 适当地选择 t 和 s，它们均可变成形式

$$y_{n+t}-2y_n+y_{n-t}=h^2\sum_{m=0}^{q}\gamma_m\nabla^m f_{n+s}. \tag{6-29}$$

由于这些公式的构造，它们对 $f(x,y(x))=y''(x)$ 是 $\leqslant q$ 次的多项式是精确的. 为了使它们对 $q+1$ 次多项式亦精确，必须在右端加上 $\gamma_{q+1}\nabla^{q+1}f_{n+s}$ 项. 由此便得 (6-29) 的阶为 p，其中 p 使 $\gamma_p\neq 0$ 的第一个 $>q$ 的整数. 为了确定误差常数，我们假设 (6-29) 为 $p=q+1$ 阶，因而 $\gamma_{q+1}\neq 0$. 若

$$y(x)=x^{q+3},$$

则由于 $\nabla^m x^m=h^m m!$，有

$$y_{n+t}-2y_n+y_{n-t}-h^2\sum_{m=0}^{q}\gamma_q\nabla^m y''_{n+s}$$

$$=h^2\gamma_{q+1}\nabla^{q+1}y''_{n+s}=h^{q+3}\gamma_{q+1}(q+3)!.$$

把上面的表达式与 (6-27) 比较，便得 $C_{p+2}=\gamma_{q+1}$. β_0+

$\beta_1+\cdots+\beta_k$ 的值当 $f(x,y)=1$ 时恒等于 $\Sigma\gamma_q\nabla^q f_{n+s}$，由此便得 $\beta_0+\beta_1+\cdots+\beta_k=\gamma_0$。 于是当 $\gamma_{q+1}\neq 0$ 时我们对所有形如 (6-29) 的方法均有

$$C=\gamma_{q+1}/\gamma_0. \tag{6-30}$$

表 6.6 给出了 §6.1-1 中讨论过的特殊方法的 p 及 C 值。

表 6.6 特殊二阶方程的特殊多步法的阶及误差常数

方　　法		有关差分算子阶数	误差常数
Störmer	$q=0$	2	$\frac{1}{12}$
	$q>0$	$k=q+1$	σ_{q+1}
Cowell	$q=2$	4	$-\frac{1}{240}$
	$q>2$	$k+1=q+1$	σ^*_{q+1}
(6-15)	$q=2$	4	$\frac{1}{60}$
	$q>2$	$k=q+1$	$\frac{1}{4}\tau_{q+1}$
(6-19)	$q=4$	6	$-\frac{1}{1890}$
	$q>4$	$k+1=q+1$	$\frac{1}{4}\tau^*_{q+1}$

q 是所采用的最高差分阶
k 是差分方程的阶.

6.1-3. 余项的界． 若 $L[y(x);h]$ 为一个 p 阶算子，我们将用 $|y^{(p+2)}(x)|$ 的界 Y 来表示 $L[y(x);h]$ 的界。 若存在广义的中值定理，使

$$L[y(x);h]=h^{p+2}C_{p+2}y^{(p+2)}(\zeta) \tag{6-31}$$

成立，其中 ζ 是某个指定区间中的点，于是立得其界为

$$|L[y(x);h]|=h^{p+2}|C_{p+2}|Y. \tag{6-32}$$

看来对任何一大类特殊算子都获得 (6-31) 是困难的。然而，对于许多有兴趣的公式，广义中值定理是可用特殊方法来证明的。在下面的讨论中，我们都假设 $y^{(p+2)}(x)$ 为连续的。

$q=0$ 时与 Störmer 方法有关的算子为

$$L[y(x);h]=y(x+h)-2y(x)+y(x-h)-h^2y''(x).$$

将 $y(x+h)$ 及 $y(x-h)$ 展成 h 的幂级数直到 h^4 项并采用带导数形式的余项，便得到

$$L[y(x);h]=\frac{1}{24}h^4\{y^{\mathrm{IV}}(\zeta_1)+y^{\mathrm{IV}}(\zeta_2)\}, \quad (6\text{-}33)$$

其中（若 $h>0$）有 $x<\zeta_1<x+h$ 及 $x-h<\zeta_2<x$. 这个公式本身就足以导出 (6-32). 为了从 (6-33) 推导出一个广义的中值定理，我们指出 $y^{\mathrm{IV}}(x)$ 作为一个连续函数在区间 $\zeta_2<x<\zeta_1$ 中取 $y^{\mathrm{IV}}(\zeta_1)$ 及 $y^{\mathrm{IV}}(\zeta_2)$ 之间的所有值. 因此，特别是对某个满足 $\zeta_2<\zeta<\zeta_1$ 的 ζ，有

$$y^{\mathrm{IV}}(\zeta)=\frac{1}{2}[y^{\mathrm{IV}}(\zeta_1)+y^{\mathrm{IV}}(\zeta_2)].$$

立刻得到关系式

$$L[y(x);h]=\frac{1}{12}h^4y^{\mathrm{IV}}(\zeta). \quad (6\text{-}34)$$

对于 Cowell 方法 $q=2$ 的情形，可用类似于证明(5-42)时所用的论据来证明，

$$y(x+h)-2y(x)+y(x-h)-\frac{1}{12}h^2\{y''(x+h)$$

$$+10y''(x)+y''(x-h)\}=\frac{-1}{240}h^6y^{\mathrm{VI}}(\zeta) \quad (6\text{-}35)$$

对于 $(x-h,x+h)$ 中的某个 ζ 是成立的.

正如逼近一阶微分方程的算子一样，每个 p 阶算子 (6-24) 可表成

$$L[y(x);h]=h^{p+2}\int_0^k G(s)y^{(p+2)}(x+sh)ds. \quad (6\text{-}36)$$

利用带有余项的积分表达式的泰劳公式，若核 $G(s)$ 在 $[0,k]$ 中不变号，我们可以推导出广义中值定理，如在第五章问题

28 中那样。若 $G(s)$ 变号，仍有

$$|L[y(x);h]| \leqslant h^{p+2}GY, \qquad (6\text{-}37)$$

其中

$$G = \int_0^k |G(s)|ds. \qquad (6\text{-}38)$$

为了导出由 (6-18) 定义的算子的广义中值定理，我们将应用表达式 (6-36)。不失一般性，令 $x_{p-1} = 0$，并在原点附近展开，得到

$$y(2h) - 2y(0) + y(-2h) - \frac{4}{3}h^2[y''(h) + y''(0) + y''(-h)] = h^6\int_{-2}^{2} G(s)y^{\mathrm{VI}}(sh)ds,$$

其中

$$G(s) = \begin{cases} \dfrac{1}{5!}(2-s)^5 - \dfrac{4}{3\cdot 3!}(1-s)^3, & 0 \leqslant s \leqslant 1, \\ \dfrac{1}{5!}(2-s)^5, & 1 \leqslant s \leqslant 2 \end{cases}$$

且 $G(-s) = G(s)$。对于 $1 \leqslant s < 2$，显然有

$$G(0) > 0, \quad G(s) > 0.$$

由关于 Descartes 符号的法则，容易证明 $G(s)$ 在 $0 < s < 1$ 中无零点。因此证明了 $-2 \leqslant s \leqslant 2$ 时 $G(s) \geqslant 0$，于是广义中值定理成立。

同样可以证明关于 (6-22) 的算子满足 $G(s) \leqslant 0$，于是广义中值定理也成立。

6.1-4. 收敛性；定义。虽然 $\beta_k \neq 0$ 时 (6-23) 为 y_{n+k} 的隐式方程，但由定理 5.4 可得，若函数 $f(x,y)$ 满足具有 Lipschitz 常数 L 的一个 Lipschitz 条件，则 (6-23) 对所有满足

$$|h| < \left|\frac{\alpha_k}{\beta_k L}\right|^{1/2} \qquad (6\text{-}39)$$

的 h 有唯一解 y_{n+k}. 这个解可用 §5.2-2 中所描述的逐次代入法来获得. 对于所有满足 (6-39) 的 h, 值

$$y_m(m = k, k+1, \cdots)$$

可视为由开始值 $y_0, y_1, \cdots, y_{k-1}$ 唯一确定的函数, 于是它们为 h 的函数:

$$y_\mu = \eta_\mu(h), \quad \mu = 0, \cdots, k-1.$$

我们期望对于一个"好"的方法来说, 假如开始值是适当选择的, 由此产生的 y_n 当 $h \to 0$ 且 $x_n = x$ 时趋向精确解在 x 点的值. 这个直观的概念可用下面收敛性定义形式化.

定义. 由 (6-23) 确定的线性多步法称为收敛的, 若下面的陈述对满足定理 1.1 条件的所有函数 $f(x,y)$ 以及一切常数 η 及 η' 正确, 即, 若 $y(x)$ 表示初值问题

$$y'' = f(x,y) \quad y(a) = \eta, \quad y'(a) = \eta' \tag{6-40}$$

的解, 则

$$\lim_{\substack{h\to 0\\ x_n = x}} y_n = y(x) \tag{6-41}$$

对所有的 $x \in [a, b]$ 以及所有由 (6-23) 确定的具有满足二个条件

$$\lim \eta_\mu(h) = \eta; \quad \mu = 0, 1, \cdots, k-1, \tag{6-42a}$$

$$\lim_{h\to 0} \frac{\eta_\mu(h) - \eta_0(h)}{\mu h} = \eta', \quad \mu = 1, \cdots, k-1 \tag{6-42b}$$

的开始值 $y_\mu = \eta_\mu(h)$ 的序列 $\{y_n\}$ 成立.

这个定义与一阶方程相应的定义其差别只是附加了条件 (6-42b), 它保证了开始值正确地近似所需要解的初始斜率. 为了使 (6-42a) 及 (6-42b) 成立, 开始值为精确的 [即 $y_\mu = y(x_\mu)$], 这显然是充分的但不是必要的.

6.1-5. 收敛性; 稳定性条件. 将差分方程 (6-23) 或差分算子 (6-24) 与多项式

$$\begin{aligned}\rho(\zeta) &= \alpha_k\zeta^k + \alpha_{k-1}\zeta^{k-1} + \cdots + \alpha_0,\\ \sigma(\zeta) &= \beta_k\zeta^k + \beta_{k-1}\zeta^{k-1} + \cdots + \beta_0\end{aligned} \tag{6-43}$$

联系在一起是方便的．反之，用任意二个具有实系数且满足 $\alpha_k \neq 0$，$|\alpha_0| + |\beta_0| \neq 0$ 的多项式(6-43)便可与一个差分算子(6-24)联系在一起．

定理 6.1. 由(6-23)确定的线性多步方法收敛的一个必要条件为多项式 $\rho(\zeta)$ 的根的模均不超过 1，且模为 1 的根其重数最多为 2．

这样加于 $\rho(\zeta)$ 的根条件称为稳定性条件．

它的证明类似于定理 5.3． 我们考虑初值问题 $y''=0$，$y(0)=y'(0)=0$，它的精确解恒等于零．若 ζ 为 $\rho(\zeta)$ 的一个根，则 $y_n = h^2\mathrm{Re}\zeta^n$ 定义了相应的差分方程的一个解，它同时满足(6-42)． 若方法收敛，便有 $\lim\limits_{n\to\infty} y_n = 0$． 对于 ζ 为实的情形，立刻得到 $|\zeta| \leqslant 1$． 若 ζ 为复，我们从二个方程

$$\frac{1}{2}h^2(\zeta^n + \bar{\zeta}^n) = y_n \text{ 及 } \frac{1}{2}h^2(\zeta^{n+1} + \bar{\zeta}^{n+1}) = y_{n+1}$$

中消去 $\bar{\zeta}^n$ 而得到

$$\frac{1}{2}h^2\zeta^n = \frac{y_n\bar{\zeta} - y_{n+1}}{\bar{\zeta} - \zeta}.$$

对于一个收敛的方法，此式右端项是收敛于零的．因此左端交收敛于零，除非 $|\zeta| \leqslant 1$，否则这是不可能的．

若 ζ 为 $\rho(\zeta)$ 的一个重数超过 2 的根，则 $y_n = h^{3/2}\mathrm{Re}\, n^2\zeta^n$ 定义差分方程的一个解．对于我们的问题，它仍满足(6-42)．因而对于一个收敛的方法来说，若 $nh = x > 0$，便有

$$\lim_{n\to\infty} y_n = 0.$$

若 ζ 为实的，立刻得到 $|\zeta| < 1$；若它为复，我们从方程

$$\frac{1}{2}(n^2h^{3/2})(\zeta^n + \bar{\zeta}^n) = y_n$$

及

$$\frac{1}{2}[(n+1)^2h^{3/2}](\zeta^{n+1}+\bar{\zeta}^{n+1})=y_{n+1}$$

中消去 ζ^n，便得

$$\zeta^n=2\frac{z_n\bar{\zeta}-z_{n+1}}{\bar{\zeta}-\zeta},$$

其中 $z_n=n^{-2}h^{-3/2}y_n(n=1,2,\cdots)$。由于当 $h\to 0$，$x_n=x>0$ 时，$z_n=h^{1/2}x^{-2}y_n\to 0$，如所希望，得到 $|\zeta|<1$。

值得注意的是 §6.1-1 中所考虑的建立在积分基础上的所有特殊方法，都满足稳定性条件。

6.1-6. 收敛性；相容性条件. 一方面稳定性条件对于小的初始扰动起着不使它过分增大的作用，但不足以保证其收敛性，还必须加上其它条件，即差分算子 (6-24) 至少局部地是微分方程的一个好的近似。

定理 6.2. 一个收敛的线性多步方法 (6-23) 的阶至少为 1。

这个相容性条件要求 §6.1-2 中所引进的常数 C_i 满足

$$C_0=C_1=C_2=0;$$

利用多项式 (6-43) 可以表示成

$$\rho(1)=0,\ \rho'(1)=0,\ \rho''(1)=2\sigma(1).\qquad(6\text{-}44)$$

注意，由于稳定性条件，对于一个收敛的方法，$\rho''(1)\neq 0$，因此 $\sigma(1)$ 也不为零。

定理 6.2 的证明. 我们将逐个证明，对于一个收敛的方法来说，由 (6-26) 定义的常数 C_0，C_1，C_2 均为零。

首先考虑初值问题 $y''=0$，$y(0)=1$，$y'(0)=0$，其精确解为 $y(x)=1$。差分方程 (6-23) 为

$$\alpha_k y_{n+k}+\alpha_{k-1}y_{n+k-1}+\cdots+\alpha_0 y_n=0,\ n=0,1,2,\cdots.\qquad(6\text{-}45)$$

假设精确开始值满足(6-42)，则对于每个 $x>0$，必有

$$\lim_{\substack{h\to 0\\ nh=x}} y_n = 1.$$

因为由(6-45)确定的 y_n 并不依赖于 h，这条件就等价于

$$\lim_{n\to\infty} y_n = 1. \tag{6-46}$$

在(6-45)中令 $n\to\infty$，便得到

$$\alpha_k + \alpha_{k-1} + \cdots + \alpha_0 = 0, \tag{6-47}$$

它正好就是 $C_0=0$.

其次，考虑问题 $y''=0$, $y(0)=0$, $y'(0)=1$，它的精确解为 $y(x)=x$. 值 y_n 仍满足(6-45). 假设

$$y_\mu = \mu h(\mu = 0,1,\cdots,k-1)$$

为精确的开始值，易见 $y_n = hz_n$，其中 $\{z_n\}$ 为(6-45)的解，而开始值为 $z_\mu=\mu$. 这个解与 h 无关. 由于方法的收敛性，对每个 $x>0$，有

$$\lim_{\substack{h\to 0\\ nh=x}} \frac{y_n}{x} = 1.$$

由此便得

$$\lim_{n\to\infty} \frac{z_n}{n} = 1. \tag{6-48}$$

将方程

$$\alpha_k z_{m+k} + \alpha_{k-1} z_{m+k-1} + \cdots + \alpha_0 z_m = 0$$

对 $m=0,1,\cdots,n$ 求和，由于(6-47)，我们得到关系式

$$\begin{aligned}\alpha_k z_{n+k} &+ (\alpha_k + \alpha_{k-1}) z_{n+k-1} + \cdots \\ &+ (\alpha_k + \alpha_{k-1} + \cdots + \alpha_1) z_{n+1} = -D,\end{aligned} \tag{6-49}$$

其中

$$\begin{aligned}D = z_0\alpha_0 &+ z_1(\alpha_1+\alpha_0) + \cdots \\ &+ z_{k-1}(\alpha_{k-1} + \alpha_{k-2} + \cdots + \alpha_0)\end{aligned}$$

是与 n 无关的量. 以 n 除以(6-49)并令 $n\to\infty$，使用(6-

48)，我们得到

$$\alpha_k + (\alpha_k + \alpha_{k-1}) + \cdots + (\alpha_k + \alpha_{k-1} + \cdots + \alpha_1) = 0$$

或

$$k\,\alpha_k + (k-1)\alpha_{k-1} + \cdots + \alpha_1 = 0.$$

它等价于 $C_1 = 0$.

最后考虑问题 $y'' = 2$, $y(0) = y'(0) = 0$, 其精确解为 $y(x) = x^2$. 差分方程 (6-23) 此时便为

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = 2h^2(\beta_k + \cdots + \beta_0). \tag{6-50}$$

我们已经证明，对于一个收敛的方法来说，$\rho(1) = \rho'(1) = 0$，并且由于稳定性，$\rho''(1) \neq 0$. 因此

$$k^2\alpha_k + (k-1)^2\alpha_{k-1} + \cdots + 1^2\alpha_1 = \rho''(1) + \rho'(1) \neq 0.$$

可以证明

$$y_n = Kh^2n^2, \tag{6-51}$$

其中

$$K = \frac{2(\beta_k + \beta_{k-1} + \cdots + \beta_0)}{k^2\alpha_k + (k-1)^2\alpha_{k-1} + \cdots + 1^2\alpha_1}$$

定义 (6-50) 的一个解，当 $\eta = \eta' = 0$，它也满足 (6-42). 因此，由于方法收敛，对于每个 $x > 0$，有

$$Kx^2 = \lim_{\substack{h\to 0\\ nh=x}} y_n = x^2.$$

由此得 $K = 1$，意思是 $C_2 = 0$. 于是便完成了定理 6.2 的证明.

易证 §6.1-1 中所描述的那些特殊方法，在定理 6.2 的意义下皆相容.

6.1-7. 最大阶算子的构造. 本节将指出：对于一个给定的满足 $\rho(1) = \rho'(1) = 0$ 的多项式 $\rho(\zeta)$，怎样将多项式 $\sigma(\zeta)$ 与它联结在一起，而使得由 (6-24) 确定的有关算子 L

具有最大可能的阶．为此目的，定义函数

$$\varphi(\zeta)=(\log\zeta)^{-2}\rho(\zeta)-\sigma(\zeta), \qquad (6-52)$$

使它在切割后的 ζ-平面取 $\log\zeta$ 的主值而成为一个单值函数．与 §5.2-7 中的证明完全相同，我们可以证明下面的引理．

引理 6.1. 与多项式 $\rho(\zeta)$ 及 $\sigma(\zeta)$ 有关的差分算子 (6-24) 确为 p 阶，其充要条件是函数 $\varphi(\zeta)$ 在 $\zeta=1$ 处有一个确为 p 次的零点。

易得与定理 5.7 类似的如下定理：

定理 6.3. 令 $\rho(\zeta)$ 为满足 $\rho(1)=\rho'(1)=0$ 的 k 次多项式．对每一个满足 $0\leqslant k'\leqslant k$ 的整数，存在唯一的一个次数 $\leqslant k'$ 的多项式 $\sigma(\zeta)$，使得与之相关联的算子 L 的阶至少为 $k'+1$．

证．由于 $\rho(\zeta)$ 在 $\zeta=1$ 处有二重根，函数 $(\log\zeta)^{-2}\rho(\zeta)$ 在 $\zeta=1$ 处解析，因而能展成一个 Taylor 级数．令

$$\frac{\rho(\zeta)}{(\log\zeta)^2}=c_0+c_1(\zeta-1)+c_2(\zeta-1)^2+\cdots, \qquad (6-53)$$

规定

$$\sigma(\zeta)=c_0+c_1(\zeta-1)+\cdots+c_{k'}(\zeta-1)^{k'}. \qquad (6-54)$$

函数 $\varphi(\zeta)$ 在 $\zeta=1$ 处就有一个重数 $\geqslant(k'+1)$ 的零点，且与之关联的算子由引理 6.1 其阶 $\geqslant(k'+1)$ (阶等于 $k'+1$ 除非 $c_{k'+1}=0$)．反之，若 L 的阶为 $k'+1$，则 $\varphi(\zeta)$ 在 $\zeta=1$ 处有一个重数为 $k'+1$ 的零点，并且由于 Taylor 展式的唯一性，仅当 $\sigma(\zeta)$ 由 (6-54) 确定时才能是这种情形．

定理 6.3 只在 $k'=k+1$ 及 $k'=k$ 时才有实际意义，前者导出最好的显式方法，而后者导出最好的隐式方法(除非发生 $c_k=0$ 的情形)．定理 6.3 的证明方法同时导出误差常数

的简单公式．若差分算子 (6-24) 有 p 阶，则在 (6-27) 中令 $y(x)=e^x$，便得到

$$\rho(e^h)-h^2\sigma(e^h)=C_{p+2}h^{p+2}+O(h^{p+3}),$$

其中 $C_{p+2}\neq 0$．将 e^h 换成 ζ 并注意当 $h\to 0$ 时

$$h=\log\zeta=\zeta-1+O((\zeta-1)^2),$$

当 $\zeta\to 1$ 时便得到，

$$(\log\zeta)^{-2}\rho(\zeta)-\sigma(\zeta)$$

$$=C_{p+2}(\zeta-1)^p+O((\zeta-1)^{p+1}). \tag{6-55}$$

由此推得，$C_{p+2}=c_p$，其中 c_p 为展开式 (6-53) 中未被吸收到 $\sigma(\zeta)$ 中去的不为零的第一项．由于 $\sigma(1)=c_0$，由此可得

$$C=c_p/c_0. \tag{6-56}$$

如果想要 (6-23) 的右端表达式用 f_{n+k} 的向后差分来表示，为了方便起见，引入变量 $t=(1-\zeta^{-1})$，那么差分 $\nabla^q f_{n+k}$ 就提供多项式 $\sigma(\zeta)$ 中 $\zeta^k t^q$ 这一项．用

$$\rho(\zeta)=\rho^k R(t),\ \sigma(\zeta)=\zeta^k S(t) \tag{6-57}$$

来确定多项式 $R(t)$ 及 $S(t)$，由于 $\log\zeta=-\log(1-t)$ 以及 $\zeta-1=t(1-t)^{-1}=t+O(t^2)$，从 (6-55) 便得到关系式

$$\frac{R(t)}{[\log(1-t)]^2}-S(t)=C_{p+2}t^p+O(t^{p+1}). \tag{6-58}$$

若 $p>k$，则可证明多项式 $S(t)$ 恒等于函数

$$[\log(1-t)]^{-2}R(t)$$

在 $t=0$ 处的 k 次 Taylor 多项式，而 C_{p+2} 为这个展开式中紧接着不为零的项的系数．

为了说明问题，令 $\rho(\zeta)=\zeta^k-2\zeta^{k-1}+\zeta^{k-2}$，我们得到 $R(t)=t^2$，且由展开式

$$\frac{t^2}{[\log(1-t)]^2}=\sigma_0^*+\sigma_1^*t+\sigma_2^*t^2+\cdots$$

(参见 §6.1-1) 便清楚地看出多项式 $S(t)$ 的系数就是 Cowell 公式 (6-10) 的系数. 从我们现在的更为一般的观点出发,采用类似的方法可以有规律地推导出在 §6.1-1 中所讨论的其它方法.

6.1-8. 稳定算子的最大阶. 前一节的构造可以应用于任何符合相容性条件的多项式 $\rho(\zeta)$, 即满足

$$\rho(1)=\rho'(1)=0.$$

通过对 $\rho(\zeta)$ 的合理选择,人们可以期望, 通过对所含参数的个数的计算, 把由此而产生的算子的阶数直推到 $2k$ 阶. 然而,当假定 $\rho(\zeta)$ 为稳定时, 并不存在这种可能性. 我们将证明,对于与这样多项式关联的算子的最高阶数,当 k 为偶数时为 $k+2$; 当 k 为奇数时为 $k+1$.

令 $\rho(\zeta)$ 为一个符合相容性且满足稳定性条件的多项式[1]. 我们再引入由 (5-111) 确定的变量 z 并由 (5-112) 确定多项式 $r(z)$ 及 $s(z)$. 由于 $\zeta=1$ 是 $\rho(\zeta)$ 的一个恰为 2 重的根, $z=0$ 是 $r(z)$的重数恰为 2 的根. 因此我们有

$$r(z)=a_2z^2+a_3z^3+\cdots+a_kz^k,$$

其中 $a_2\neq 0$. 不失一般性,可以假设

$$a_2>0. \tag{6-59a}$$

考虑到 $r(z)$ 的乘积形式并利用 $r(z)$ 无一根具有正的实部的事实,正如 §5.2-8 中所叙述的可得

$$a_\mu\geqslant 0,\quad \mu=3,\cdots,k. \tag{6-59b}$$

现在我们考虑函数

$$p(z)=\left(\frac{1-z}{2}\right)^k\varphi\left(\frac{1+z}{1-z}\right)=\left[\log\frac{1+z}{1-z}\right]^{-2}r(z)$$
$$-s(z). \tag{6-60}$$

---

1) 我们真正需要的只是下面的较弱形式的稳定性条件: $\rho(\zeta)$ 无一根按模大于 1, 且在 $\zeta=1$ 恰为二重根.

由于$\zeta-1=2z+O(z^2)$，$p(z)$在$z=0$处有一个零点当且仅当$\varphi(\zeta)$在$\zeta=1$处有一个同阶的零点，因此，由引理 6.1，当且仅当有关的算子为p阶。令

$$\left[\frac{z}{\log\frac{1+z}{1-z}}\right]^2\frac{r(z)}{z^2}=b_0+b_1z+b_2z^2+\cdots,\tag{6-61}$$

由此得出，若该算子为p阶，则

$$s(z)=b_0+b_1z+\cdots+b_{p-1}z^{p-1}.\tag{6-62}$$

另一方面，$s(z)$最多为k次。因此，有关的算子的阶超过$k+1$当且仅当$b_{k+1}=\cdots=b_{p-1}=0$。令

$$\left[\frac{z}{\log\frac{1+z}{1-z}}\right]^2=d_0+d_2z^2+d_4z^4+\cdots,\tag{6-63}$$

当$\nu>k$时规定$a_\nu=0$，由(6-61)便得到

$$\begin{aligned}&b_0=d_0a_2,\\&b_1=d_0a_3,\\&b_{2\nu}=d_0a_{2\nu+2}+d_2a_{2\nu}+\cdots+d_{2\nu}a_2,\\&b_{2\nu+1}=d_0a_{2\nu+3}+d_2a_{2\nu+1}+\cdots+d_{2\nu}a_3,\\&\qquad\qquad\nu=1,2,\cdots.\end{aligned}\tag{6-64}$$

下面我们将证明

$$d_{2\nu}<0,\quad \nu=1,2,\cdots.\tag{6-65}$$

假设这个不等式正确，我们区分两种情形。

(i) 若k为奇数，利用(6-65)及(6-59)，则由(6-64)便得到

$$b_{k+1}=d_2a_{k+1}+d_4a_{k-1}+\cdots+d_{k+1}a_2<0,$$

而且$p>k+1$是不可能的。因此有

定理 6.4. 其步数k为奇数的稳定算子的阶不能超过$k+1$.

$k+1$阶算子的存在性已经由定理 6.3 暗示。

(ii) 若 k 为偶数，则有

$$b_{k+1} = d_4 a_{k-1} + d_6 a_{k-3} + \cdots + d_k a_3.$$

由于 (6-59) 及 (6-65)，$b_{k+1} = 0$ 成立的充要条件为

$$a_3 = a_5 = \cdots = a_{k-1} = 0.$$

这种情形出现当且仅当 $r(z)$ 为偶函数，$r(z) = r(-z)$. 由于在稳定性假设下，$r(z)$ 不能有任何具有正实部的根，故并不能有任何负实部的根. 由此得出 $r(z)$ 的所有的根均为纯虚数，且 $\rho(\zeta)$ 的所有的根其模为 1. 由于 (6-59) 及 (6-65)，

$$b_{k+2} = d_4 a_k + d_6 a_{k-2} + \cdots + d_{k+2} a_2 < 0,$$

所以阶不能超过 $k+2$. 因此我们便证明了：

定理 6.5. 一个稳定算子的阶 p 不能超过 $k+2$. 对于 $p = k+2$ 的充要条件是 k 为偶数，$\rho(\zeta)$ 的所有根其模为 1，并且 $\sigma(\zeta)$ 由 (6-54) 所确定.

我们称一个满足定理 6.5 的条件的算子为最佳算子. 定理 (6.5) 的证明方法同时提供了某些表达式以及关于一个最佳算子误差常数的不等式. 由于 $\zeta - 1 = 2z + O(z^2)$，从 (6-55) 得到

$$p(z) = 2^{p-k} C_{p+2} z^p + O(z^{p+1}).$$

若 $p > k$，因此便有 $C_{p+2} = 2^{k-p} b_p$. 利用

$$\sigma(1) = 2^k s(0) = 2^k b_0,$$

我们得到

$$C = b_p / 2^p b_0. \tag{6-66}$$

对于一个最佳算子，$p = k+2$. 因此，利用 (6-64)，便有

$$C = \frac{d_4 a_k + d_6 a_{k-2} + \cdots + d_{2k+2} a_2}{2^{k+2} d_0 a_2}. \tag{6-67}$$

由于 (6-59) 及 (6-65)，并利用数值 $d_0 = \frac{1}{4}$，$d_4 = -\frac{1}{60}$，我们很容易得到以下的关于 C 的不等式：

$$C \leqslant -\frac{1}{2^k 60} \frac{a_k}{a_2}, \tag{6-68}$$

$$C \leqslant 2^{-k} d_{k+2}. \tag{6-69}$$

对于 Cowell 方法(6-14)，二种情形均有等式成立。

为了求得系数 $d_{2\nu}$ 的递推关系，同时也为了证明(6-65)，我们将计算以下展开式中的系数 A_ν，

$$\left[\frac{1}{2}\log\frac{1+z}{1-z}\right]^2 = A_0 z^2 + A_1 z^4 + A_2 z^6 + \cdots.$$

对两边微分，便得到

$$2A_0 z + 4A_1 z^3 + 6A_2 z^5 + \cdots = \frac{1}{1-z^2}\log\frac{1+z}{1-z}$$

$$= 2(1 + z^2 + z^4 + \cdots)\left(z + \frac{1}{3}z^3 + \frac{1}{5}z^5 + \cdots\right).$$

比较两边 $z^{2\nu+1}$ 的系数，很快便得到

$$A_\nu = \frac{1}{\nu+1} k_\nu, \quad \nu = 0, 1, 2, \cdots, \tag{6-70}$$

其中

$$k_\nu = 1 + \frac{1}{3} + \cdots + \frac{1}{2\nu+1}.$$

从

$$(d_0 + d_2 z^2 + \cdots) z^{-2}\left[\log\frac{1+z}{1-z}\right]^2 = 1,$$

我们便有递推关系 $d_0 = \frac{1}{4}$，

$$d_{2\nu} = -\frac{1}{2}k_1 d_{2\nu-2} - \frac{1}{3}k_2 d_{2\nu-4} - \cdots - \frac{1}{\nu+1}k_\nu d_0,$$

$$\nu = 1, 2, \cdots,$$

由此便可很快地计算出表 6.7 中的数值。

为了证明(6-65)，我们再应用 Kaluza 的引理 5.4 于函

表 6.7 (6-63)中系数 $d_{2\nu}$ 的值

ν	0	1	2	3	4
$d_{2\nu}$	$\frac{1}{4}$	$-\frac{1}{6}$	$-\frac{1}{60}$	$-\frac{8}{945}$	$-\frac{76}{14175}$

数

$$f(t) = \left[\frac{1}{2z}\log\frac{1+z}{1-z}\right]^2 = A_0 + A_1 t + A_2 t^2 + \cdots,$$

其中 $t = z^2$. 显然 $A_\nu > 0$，为了证明

$$\Delta_\nu = A_{\nu+1}A_{\nu-1} - A_\nu^2 > 0,$$

我们注意到，利用 (6-70)，对 $\nu \geqslant 2$，有

$$\nu^2(\nu^2-1)\Delta_{\nu-1}^2 = k_{\nu-1}^2 - \frac{\nu^2-1}{4\nu^2-1}(2k_{\nu-1}-1)$$

$$\geqslant k_{\nu-1}^2 - \frac{1}{2}k_{\nu-1} + \frac{4}{15}.$$

最后的表达式当 $k_\nu \geqslant 1$ 时为正，所以 $\Delta_{\nu-1} > 0$. 因此便满足 Kaluza 的引理的假定，于是结论对 $C_\nu = 4d_{2\nu}(\nu = 1, 2, \cdots)$ 成立.

有趣的是：指数 >2 时，类似于 (6-65) 的关系不成立，因而本节的结果不能直接推广到当 $\mu > 2$ 时具有形式

$$y^{(\mu)} = f(x, y)$$

的微分方程的算子.

6.1-9. 最佳算子的构造. 与一个最佳方法有关的多项式 $r(z)$ 为偶这一点已经说明. 由 (6-61) 就得出 $s(z)$ 也为一偶多项式. 从 (5-112) 得到

$$\rho(\zeta) = \zeta^k\rho(\zeta^{-1}), \quad \sigma(\zeta) = \zeta^k\sigma(\zeta^{-1}),$$

或

$$\alpha_{k-\nu} = \alpha_\nu, \quad \beta_{k-\nu} = \beta_\nu, \quad \nu = 0, 1, \cdots, k.$$

从这里得出 $\rho(\zeta)$ 及 $\sigma(\zeta)$ 可以写成 $\zeta^{\frac{1}{2}k}$ 乘以某个变量为

$$\zeta+\zeta^{-1}-2$$

的 $\frac{1}{2}k$ 次多项式．此外，由于 $\zeta=1$ 为 $\rho(\zeta)$ 的一个重根，这表示 $\rho(\zeta)$ 的多项式含有一个 $\zeta+\zeta^{-1}-2$ 的因子．令

$$\zeta+\zeta^{-1}-2=t^2,$$

因而能分别找到二个 $\frac{1}{2}k-1$ 及 $\frac{1}{2}k$ 次的多项式 P 及 Q，使得

$$\rho(\zeta)=\zeta^{\frac{1}{2}k}t^2P(t^2),\quad \sigma(\zeta)=\zeta^{\frac{1}{2}k}Q(t^2).$$

由于

$$\zeta=\left(\frac{1}{2}t+\sqrt{1+\frac{1}{4}t^2}\right)^2,$$

这里的平方根的定义与 §5.2-9 中相同，我们有

$$\zeta-1=t+O(t^2)$$

及

$$\log\zeta=2\log\left(\frac{1}{2}t+\sqrt{1+\frac{1}{4}t^2}\right).$$

于是关系式 (6-55) 等价于

$$\left\{\frac{\frac{1}{2}t}{\log\left(\frac{1}{2}t+\sqrt{1+\frac{1}{4}t^2}\right)}\right\}^2P(t^2)-Q(t^2)$$
$$=C_{k+4}t^{k+2}+O(t^{k+4}).$$

左端第一项表示 $t=0$ 处解析的偶函数．令

$$\left\{\frac{\frac{1}{2}t}{\log\left(\frac{1}{2}t+\sqrt{1+\frac{1}{4}t^2}\right)}\right\}^2P(t^2)$$
$$=q_0+q_2t^2+q_4t^4+\cdots,\qquad(6\text{-}71)$$

由 Taylor 展式的唯一性便得到

$$Q(t^2) = q_0 + q_2 t^2 + \cdots + q_k t^k. \qquad (6\text{-}72)$$

此外，由于 $\sigma(1) = Q(0) = q_0$，便有

$$C = q_{k+2}/q_0. \qquad (6\text{-}73)$$

将多项式

$$P(t^2) = p_0 + p_2 t^2 + \cdots + p_{k-2} t^{k-2}$$

乘以 Taylor 级数

$$\left\{\frac{\frac{1}{2}t}{\log\left(\frac{1}{2}t + \sqrt{1 + \frac{1}{4}t^2}\right)}\right\}^2 = l_0 + l_2 t^2 + l_4 t^4 + \cdots, \qquad (6\text{-}74)$$

便可计算系数 $q_{2\nu}$。从

$$\frac{\log\left(\frac{1}{2}t + \sqrt{1 + \frac{1}{4}t^2}\right)}{\frac{1}{2}t} = 1 + \frac{1}{3}\binom{-\frac{1}{2}}{1}\left(\frac{t}{2}\right)^2 + \frac{1}{5}\binom{-\frac{1}{2}}{2}\left(\frac{t}{2}\right)^4 + \cdots$$

可得系数 $l_{2\nu}$ 满足递推关系式 $l_0 = 1$，

$$n l_{2n} = -\frac{n+1}{3\cdot 2^2}\binom{-\frac{1}{2}}{1} l_{2n-2} - \frac{n+2}{5\cdot 2^4}\binom{-\frac{1}{2}}{2} l_{2n-4} - \cdots - \frac{2n}{(2n+3)\cdot 2^{2n}}\binom{-\frac{1}{2}}{n} l_0,$$

$$n = 1, 2, \cdots.$$

现在易得表 6.8 中的数值.

表 6.8 (6-74) 中系数 $l_{2\nu}$

ν	0	1	2	3	4
$l_{2\nu}$	1	$\frac{1}{12}$	$-\frac{1}{240}$	$\frac{31}{60480}$	$\frac{289}{3628800}$

例. (i) 当 $k=2$ 时的最佳方法. 我们有 $P(t^2)=1$, 因此便有 $q_{2\nu}=2^{-2\nu}l_{2\nu}$, $Q(t^2)=1+\frac{1}{12}t^2$. 由此而产生的 Cowell 公式 (6-14) 及误差常数值为 $C=l_4=-\frac{1}{240}$, 这与前面的计算相符.

(ii) 当 $k=4$ 时一般的最佳方法. 假设 $\rho(\zeta)$ 的非零根在点 $\zeta=e^{\pm i\varphi}$ 处,便有

$$\rho(\zeta)=\zeta^2\left(\zeta+\frac{1}{\zeta}-2\right)\left(\zeta+\frac{1}{\zeta}-2\cos\varphi\right). \quad (6-75)$$

因此 $P(t^2)=4\lambda+t^2$, 其中 $\lambda=\sin^2\frac{1}{2}\varphi$. 因而便得到

$$\left\{\frac{\frac{1}{2}t}{\log\left(\frac{1}{2}t+\sqrt{1+\frac{1}{4}t^2}\right)}\right\}^2 P(t^2)=4\lambda+\left(1+\frac{\lambda}{3}\right)t^2$$

$$+\left(\frac{1}{12}-\frac{\lambda}{60}\right)t^4+\left(-\frac{1}{240}+\frac{31\lambda}{15120}\right)t^6+\cdots.$$

于是便有

$$Q(t^2)=4\lambda+\left(1+\frac{\lambda}{3}\right)t^2+\left(\frac{1}{12}-\frac{\lambda}{60}\right)t^4,$$

$$C=\frac{31}{15120}-\frac{1}{960\lambda}.$$

当 $\lambda=1$, $\varphi=\pi$ 时, 又得到 (6-22). 当 $\lambda=\frac{3}{4}$ 及 $\lambda=\frac{1}{2}$ 时,我们得到新的公式

$$y_{n+4}-y_{n+3}-y_{n+1}+y_n=h^2\left\{3f_{n+2}+\frac{5}{4}\nabla^2 f_{n+3}+\frac{17}{24}\nabla^4 f_{n+4}\right\}. \tag{6-76}$$

$$y_{n+4}-2y_{n+3}+2y_{n+2}-2y_{n+1}+y_n$$
$$=h^2\left\{2f_{n+2}+\frac{7}{6}\nabla^2 f_{n+3}+\frac{3}{40}\nabla^4 f_{n+4}\right\}. \tag{6-77}$$

当 $\lambda=1$ 时，误差常数达到上界 (6-69)。

(iii) $k=6$ 时的一个特殊方法．多项式

$$\rho(\zeta)=\zeta^6-\zeta^4-\zeta^2+1$$

的非平凡根为 $-1, -1, +i, -i$，因此满足定理 6.5 的条件，有

$$\rho(\zeta)=\zeta^3(\zeta+\zeta^{-1}-2)(\zeta+\zeta^{-1})(\zeta+\zeta^{-1}+2),$$

因而

$$P(t^2)=8+6t^2+t^4.$$

我们的算法给出

$$Q(t^2)=8+\frac{20}{3}t^2+\frac{22}{15}t^4+\frac{59}{945}t^6,$$

因此公式升高到 8 阶：

$$y_{n+6}-y_{n+4}-y_{n+2}+y_n=h^2\left\{8f_{n+3}+\frac{20}{3}\nabla^2 f_{n+4}\right.$$
$$\left.+\frac{22}{15}\nabla^4 f_{n+5}+\frac{59}{945}\nabla^6 f_{n+6}\right\}. \tag{6-78}$$

6.2. 离散误差

6.2-1．两个引理．本节将证明类似于引理 5.5 及 5.6 的两个引理．第一个引理为证明第二个所需要，而第二个引理则为获得下节将推导的离散误差先验界的主要工具．

引理 6.2. 令多项式 $\rho(\zeta)=\alpha_k\zeta^k+\cdots+\alpha_0$ 满足二阶

方程积分的稳定性条件，且令系数 $\gamma_l(l=0,1,2,\cdots)$ 由

$$\frac{1}{\alpha_k+\alpha_{k-1}\zeta+\cdots+\alpha_0\zeta^k}=\gamma_0+\gamma_1\zeta+\gamma_2\zeta_2+\cdots \tag{6-79}$$

来确定，则存在二个常数 Γ 及 γ，使得

$$|\gamma_l|\leqslant l\Gamma+\gamma,\quad l=0,1,2,\cdots \tag{6-80}$$

成立。

证。 令 $\alpha_k+\alpha_{k-1}\zeta+\cdots+\alpha_0\zeta^k=\hat{\rho}(\zeta)$，便有 $\hat{\rho}(\zeta)=\zeta^k\rho(\zeta^{-1})$. 由于 $\rho(\zeta)$ 在 $|\zeta|=1$ 之外无根，且由于在 $|\zeta|=1$ 上的根重数至多为 2，多项式 $\hat{\rho}(\zeta)$ 在 $|\zeta|=1$ 内无根，且它在 $|\zeta|=1$ 上的根重数最多为 2。

若把 $\rho(\zeta)$ 在 $|\zeta|=1$ 上的重根记为 $\zeta_2,\zeta_4,\cdots,\zeta_{2d}$，则对于适当选择的常数 $A_1,\cdots,A_d$，函数

$$f(\zeta)=\frac{1}{\hat{\rho}(\zeta)}-\frac{A_1}{(\zeta-\zeta_2^{-1})^2}-\cdots-\frac{A_d}{(\zeta-\zeta_{2d}^{-1})^2} \tag{6-81}$$

当 $|\zeta|<1$ 时为全纯且在 $|\zeta|=1$ 上至多有有限个单极点。由引理 5.5，它的 Taylor 展式在 $\zeta=0$ 处的系数是有界的。由于

$$\frac{1}{(\zeta-\zeta_\mu^{-1})^2}=\frac{\zeta_\mu^2}{(1-\zeta_\mu\zeta)^2}$$

$$=\zeta_\mu^2\{1+2\zeta_\mu\zeta+3(\zeta_\mu\zeta)^2,+\cdots\}, \tag{6-82}$$

(6-81) 右边各项在 $\zeta=0$ 处的 Taylor 系数满足形如 (6-80) 的不等式. 由此便得出对于函数

$$\frac{1}{\hat{\rho}(\zeta)}=f(\zeta)+\sum_{\mu=1}^{d}\frac{A_\mu}{(\zeta-\zeta_{2\mu}^{-1})^2} \tag{6-83}$$

有同样的估计式成立. 下面的例子说明在具体情形下易得 Γ

及 γ 值. 对于 Störmer 及 Cowell 方法,我们有

$$\rho(\zeta) = \zeta^k(1-\zeta^{-1})^2,$$

及 $\hat{\rho}(\zeta) = (1-\zeta)^2$. 由 (6-82) 得 (6-80) 是正确的, 对 $\Gamma = 1$, $\gamma = 1$. 对于"双步"方法 (6-15) 及 (6-19), 有 $\rho(\zeta) = \zeta^k(1-\zeta^{-2})^2$ 及 $\hat{\rho}(\zeta) = (1-\zeta^2)^2$. 由此取 $\Gamma = \frac{1}{2}$, $\gamma = 1$, (6-80) 成立.

下面的引理与差分方程:

$$\begin{aligned} &\alpha_k z_{m+k} + \alpha_{k-1} z_{m+k-1} + \cdots + \alpha_0 z_m \\ &\quad = h^2\{\beta_{k,m} z_{m+k} + \cdots + \beta_{0,m} z_m\} + \lambda_m \end{aligned} \tag{6-84}$$

解的增长有关.

引理 6.3. 令多项式 $\rho(\zeta) = \alpha_k\zeta^k + \cdots + \alpha_0$ 满足稳定性条件(关于二阶方程积分),令 B^*, β 及 Λ 为使

$$|\beta_{k,m}| + |\beta_{k-1,m}| + \cdots + |\beta_{0,m}| \leqslant B^*, \quad |\beta_{k,m}| \leqslant \beta$$

$$|\lambda_m| \leqslant \Lambda, \quad 0 \leqslant m \leqslant N \tag{6-85}$$

成立的常数, 并令 $0 \leqslant h^2 < |\alpha_k|\beta^{-1}$, 那么 (6-84) 的每一个使

$$|z_\mu| \leqslant Z, \quad \mu = 0, 1, \cdots, k-1 \tag{6-86}$$

成立的解满足

$$|z_n| \leqslant K^* e^{nh^2L^*}, \quad 0 \leqslant n \leqslant N, \tag{6-87}$$

这里

$$L^* = \frac{(N\Gamma + \gamma)B^*}{1 - h^2|\alpha_k^{-1}|\beta},$$

$$K^* = \frac{\left(\frac{1}{2}N^2\Gamma + N\gamma\right)\Lambda + (N\Gamma + \gamma)KAz}{1 - h^2|\alpha_k^{-1}|\beta}, \tag{6-88}$$

其中 $A = |\alpha_k| + |\alpha_{k-1}| + \cdots + |\alpha_0|$, 而 Γ 与 γ 是由(6-80)确定的.

这个证明类似于引理 5.6 的证明,对 $l = 0, 1, \cdots, n-k$,

在对应于 $m=n-k-l$ 的差分方程 (6-84) 上乘以 γ_l 并将所得的方程相加，在左边我们得到 z_n 加上某些余下的由

$$m=0,1,\cdots,k-1$$

而来的项，它们由 (6-86) 便得其界为 $KAZ(N\Gamma+\gamma)$. 右边有由 $l=0$ 所提供的项 $\beta_{k,n-k}\gamma_0 z_n$，一个以

$$(N\Gamma+\gamma)B\sum_{m=0}^{k-1}|z_m|$$

为界的和式以及 λ_m 的值的加权和，对于这点由 (6-80) 及(6-85) 可得其估计式为

$$\left|\sum_{l=0}^{n-k}\gamma_l\lambda_{n-k-l}\right|\leqslant\Lambda\sum_{l=0}^{n-k}(l\Gamma+\gamma)\leqslant\Lambda\left(\frac{1}{2}N^2\Gamma+N\gamma\right).$$

从所得到的不等式中解出 $|z_n|$，便得到

$$|z_n|\leqslant h^2L^*\sum_{m=0}^{n-1}|z_m|+K^*.$$

现在我们由归纳法可得如 §5.3-2 中那样的估计式

$$|z_n|\leqslant K^*(1+h^2L^*)^n,$$

由它易得结论 (6-87).

6.2-2. **收敛性的充分条件；一个先验界.** 如同一阶方程的多步方法的情形，早已知道稳定性与相容性条件是收敛的必要条件，也是收敛的充分条件.

定理 6.6. 由 (6-23) 确定的线性多步方法是收敛的，当且仅当它满足稳定性和相容性条件.

结论中的“仅当”部分是定理 6.1 与 6.2 已经证明了的内容. “当”这一部分的证明可以建立在引理 6.3 的基础上，正如定理 5.10 的证明建立在引理 5.6 上一样，细节从略.

现在我们将在精确解 $y(x)$ 在 $x\in[a,b]$ 中具有 $p+2$ 阶连续导数的假设下讨论离散误差的界. 令

$$Y=\max_{a\leqslant x\leqslant b}|y^{(p+2)}(x)|, \tag{6-89}$$

只假设数值解 y_n 满足关系式

$$\alpha_k y_{m+k} + \cdots + \alpha_0 y_m = h^2(\beta_k f_{m+k} + \cdots + \beta_0 f_m) + \theta_m K h^{q+2}, \qquad (6\text{-}90)$$

其中 $|\theta_m| \leqslant 1$，而K及 q 都是常数，$q > 0$. 引入含有K的项是为了允许小的局部误差，同时也因为以后研究数值解的渐近性态时需要. 对于开始值误差，假设

$$|y_\mu - y(x_\mu)| \leqslant h\delta, \quad \mu = 0, 1, \cdots, k-1. \qquad (6\text{-}91)$$

同时令

$$A = |\alpha_k| + \cdots + |\alpha_0|, B = |\beta_k| + \cdots + |\beta_0|,$$

并以 Γ 和 γ 表示曾在 §6.2-1 中引入的常数

$$\Gamma^* = \frac{\Gamma}{1 - h^2 L|\alpha_k^{-1}\beta_k|}, \quad a^* = a - \frac{h\gamma}{\Gamma}.$$

最后，我们将用G来表示由 (6-38) 确定的常数. 于是我们便能证明:

定理 6.7. 在上述条件下，如果 $h^2 < L^{-1}|\alpha_k \beta_k^{-1}|$，那么离散误差 $e_n = y_n - y(x_n)$，$a \leqslant x_n \leqslant b$，满足

$$|e_n| \leqslant \Gamma^* \left[(x_n - a^*)kA\,\delta + \frac{(x_n - a^*)^2}{2}(Kh^q + GYh^p)\right] \times \exp[(x_n - a^*)^2 \Gamma^* L B]. \qquad (6\text{-}92)$$

证. 将精确解 $y(x)$ 代入差分算子 (6-24) 中，便得到

$$\alpha_k y(x_n + kh) + \cdots + \alpha_0 y(x_m) = h^2\{\beta_k y''(x_m + kh) + \cdots + \beta_0 y''(x_m)\} + R_m, \qquad (6\text{-}93)$$

其中，由 §6.1-2 的结果有

$$|R_m| \leqslant GYh^{p+2}.$$

通过

$$g_m = \begin{cases} \dfrac{f(x_m, y_m) - f(x_m, y(x_m))}{y_m - y(x_m)}, & \text{如果 } y_m \neq y(x_m), \\ 0, & \text{如果 } y_m = y(x_m) \end{cases}$$

来定义数 $g_m (m = 0, 1, 2, \cdots)$. 从 (6-90) 中减去 (6-93)，

我们得到

$$\alpha_k e_{m+k} + \cdots + \alpha_0 e_m = h^2\{\beta_k g_{m+k} e_{m+k} + \cdots + \beta_0 g_m e_m\} + \theta'_m(Kh^{q+2} + GYh^{p+2}).$$

对这个差分方程应用引理 6.3，令 $z_m = e_m$，$Z = h\delta$，$N = (x_n - a)/h$，$\Lambda = Kh^{q+2} + GYh^{p+2}$，$\beta_{\mu,m} = g_{m+\mu}\beta_\mu$. 由于 Lipschitz 条件，从 $|g_m| \leqslant L$ 我们得到

$$B^* = LB, \ \beta = L|\beta_k|.$$

利用关系式 (6-87) 以及关系式

$$N\Gamma + \gamma = (x_n - a^*)\Gamma h^{-1},$$

$$\frac{1}{2}N^2\Gamma + N\gamma = \frac{1}{2}(x_n - a)^2\Gamma h^{-2} + (x_n - a)\gamma h^{-1}$$

$$\leqslant \frac{1}{2}(x_n - a^*)^2\Gamma h^{-2},$$

便得到定理中的不等式． 正如在一阶情形中那样 [参看 (5-180)]，估计式 (6-92) 十分清楚地显示了开始误差局部不精确性，以及局部离散误差对累积误差的影响．对于每种情形，这个影响均比一阶的情形加强了一个 $1/h$ 的因子．

要求读者对某些具体特殊方法写出估计式 (6-92)，明确地以数值表示它对 Y 及 L 的依赖性．

6.2-3． **离散误差的渐近公式．** 为了研究误差当 $h \to 0$ 时的渐近性态，我们将假设除在 §6.2-2 开始时说明的那些条件外，附加下面的条件：

(i) $y^{(p+3)}(x)$ 在 $[a,b]$ 上连续；

(ii) 函数 $g(x) = f_y(x, y(x))$ 在 $[a,b]$ 上连续可微． 为了对每个 h 值确定数值解，开始值必须说明为 h 的函数．我们将记

$$e_\mu = \delta_\mu(h), \ \mu = 0, 1, \cdots, k-1, \tag{6-94}$$

并假设函数 $\delta_\mu(h)$ 为

(iii) 在 $h=0$ 处 p 次可微，且

(iv) 其阶为 $O(h^{q+1})$，其中 q 为一正常数。

在上面的假设下，我们将研究差分方程 (6-23) 即在 (6-90) 中 $K=0$ 的精确解的性态。为了将结果公式化，有必要引进一些记号。

令 $\zeta_1,\zeta_2,\cdots,\zeta_k$ 表示多项式 $\rho(\zeta)$ 的根，假设有 $2d$ 个模为 1 的重根（每个算二次），我们将这些根记为 $\zeta_1=\zeta_2=1$，$\zeta_3=\zeta_4,\cdots,\zeta_{2d-1}=\zeta_{2d}$。这些根被称为本性根。根 ζ_ν 对于 $2d<\nu\leqslant k$（如果有的话）不是单根就是模 <1 的根。对于 $\nu=1,2,\cdots,2d$，我们定义多项式

$$\rho_\nu(\zeta)=\frac{\rho(\zeta)}{\zeta-\zeta_\nu}=\alpha_{\nu,0}+\alpha_{\nu,1}\zeta+\cdots+\alpha_{\nu,k-1}\zeta^{k-1}, \tag{6-95}$$

如果 $r=\min(p,q)$ 并令

$$\Delta_\nu=\lim_{h\to 0}h^{-r-1}\{\alpha_{\nu,0}\delta_0(h)+\alpha_{\nu,1}\delta_1(h)+\cdots+\alpha_{\nu,k-1}\delta_{k-1}(h)\}, \tag{6-96}$$

显然，如果 $q>p$，则 $\Delta_\nu=0$。

函数 $e(x)$ 定义为初值问题：

$$\begin{aligned}&e''(x)=g(x)e(x)-Cy^{(p+2)}(x)\\&e(a)=e'(a)=0\end{aligned} \tag{6-97}$$

的解，其中 C 为方法的误差常数，正象由 (6-28) 确定的那样。我们将每个本性根与增长参数

$$\mu_\nu^2=\frac{2\sigma(\zeta_\nu)}{\zeta_\nu^2\rho''(\zeta_\nu)},\quad \nu=1,2,\cdots,2d \tag{6-98}$$

联系在一起。函数 $s_k(x)$ 当 $k=1,2,\cdots,d$ 时定义为初值问题：

$$\begin{aligned}&s_k''(x)=\mu_{2k}^2g(x)s_k(x),\\&s_k(a)=0,\quad s_k'(a)=1\end{aligned} \tag{6-99}$$

的解．用这些记号及定义，我们便有下面的定理．

定理 6.8. 在 §6.2-3 开始所叙述的条件下，由方法 (6-23) 定义的初值问题 (6-40) 解的离散误差，当 $h\to 0$，$nh=x-a$，$a<x\leqslant b$ 时它的渐近性态由公式

$$e_n=h^p c(x)+h^r\sum_{k=1}^{d}\frac{2\Delta_{2k}}{\rho''(\zeta_{2k})}\zeta_{2k}^{n}s_k(x)+O(h^{r+1}) \tag{6-100}$$

所描述．它的证明类似于定理 5.12 的证明，故省略．

若 $p\neq q$，(6-100) 的右边仅有一项是有意义的．若

$$p=q=r,$$

我们可以区分真的离散误差 [以包含 $c(x)$ 的项所表示的] 与开始误差 [以包含函数 $s_k(x)$ 的项所表示的]． 若 $g(x)<0$，函数 $c(x)$ 通常当 $x\to\infty$ 时至多像 x 一样地增长．另一方面，对于某个 k，若 $\mu_{2k}^2<0$，相应的函数 $s_k(x)$ 便指数地增长．这是不规则误差增长的另一例，它被称为条件稳定性．利用相容性，由 $\mu_2^2=1$，在这种意义下的条件稳定性只能由不同于 1 的重本性根所引起．现在稍许再留意一下 §6.1-7 便发现，与一阶情形相反，存在着具有最高阶 $k+2$ 的方法，对它来说 $\zeta=1$ 为它的仅有的本性根，故不存在条件稳定的问题．

然而本性根 $\zeta=1$ 可以引起不同类型的不稳定性．为了具体化的缘故，考虑初值问题

$$y''=y,\ y(0)=1,\ y'(0)=-1.$$

这个问题具有指数下降的解 $y(x)=e^{-x}$． 另一方面，由于 $s_1(x)=\sinh x$，开始误差包括着指数增长的分量． 然而这种类型的不稳定性并不是因为方法的不完美，而是已经固有在数学问题本身中的．如果对于同样的微分方程而将初始条件改变为 $y(0)=1+2\delta$，$y'(0)=-1$，其中 δ 为任意的小量，解便改变成 $y(x)=(1+\delta)e^{-x}+\delta e^{x}$，当 $x\to\infty$ 时它趋于

无穷. 精确解的这个特征使初始条件进行任意微小的改变时便使解很快地变化，而在这种情形下数值方法无非是正确地反应了数学问题的这一性质. 因此所述的这种现象被称为数学的不稳定性[1].

在定理 6.8 的基础上，人们便可以着手去导出与 §5.3-7 及 5.3-8 中完全类似的那些结论. 由于在这个分析过程中并不出现新的观点，因而将细节留给读者.

6.3. 舍入误差的传播

6.3-1. 一个先验界. 按照我们所建立的方法，将数值近似值 $\{\tilde{y}_n\}$ 所满足的方程写成如下形式:

$$\begin{aligned}\alpha_k\tilde{y}_{n+k}+\alpha_{k-1}\tilde{y}_{n+k-1}+\cdots+\alpha_0\tilde{y}_n\\ =h^2\{\beta_kf(x_{n+k},\tilde{y}_{n+k})+\cdots+\beta_0f(x_n,\tilde{y}_n)\}\\ +\varepsilon_{n+k},n=0,1,2,\cdots.\end{aligned}\tag{6-101}$$

局部舍入误差 ε_n 依赖于计算过程和计算设备的运算部件的组成情况. 本节将在对 ε_n 逐渐加严的假设下推导出关于(累积)舍入误差的结果. 开始我们仅假设 $|\varepsilon_n|\leqslant\varepsilon$，其中 ε 是与 n 无关的量. 由 (6-101) 减去相应的方程 (6-23)，并令

$$\begin{aligned}&g_m=r_m^{-1}[f(x_m,\tilde{y}_m)-f(x_m,y_m)],\quad r_m\neq0,\\ &g_m=0,r_m=0,\end{aligned}$$

便得到

$$\begin{aligned}\alpha_kr_{n+k}+\cdots+\alpha_0r_n=h^2\{\beta_kg_{n+k}r_{n+k}+\cdots+\beta_0g_nr_n\}\\ +\varepsilon_{n+k}.\end{aligned}\tag{6-102}$$

应用引理 6.3 于这个关系式，令 $z_m=r_m$, $Z=0$(没有初始舍入误差), $\Lambda=\varepsilon$, $N=(x_n-a)/h$, $B^*=L\sum|\beta_\mu|$, $\beta=L|\beta_k|$.

1) 如何有效地处理数学不稳定题目的问题，为 Fox 及 Mitchell [1957] 所考虑.

假设 $h^2 < L^{-1}|\beta_k^{-1}\alpha_k|$，当 $a \leqslant x_n \leqslant b$ 时，得到

$$|r_n| \leqslant \varepsilon h^{-2}\Gamma^*(x_n - a^*)^2 \exp[(x_n - a^*)^2\Gamma^* L B], \tag{6-103}$$

其中常数 Γ^*，a^* 及 B 像在定理 6.7 中那样确定。

(6-103) 的明显性质包含了这样事实，即在 $|r_n|$ 的界中现在以 $O(\varepsilon/h^2)$ 代替了到目前为止积分一阶方程时所有方法所获得的 $O(\varepsilon/h)$。虽然 (6-103) 仅表示一个非常保守的界，不能肯定何时达到，而下面的研究将证明方法 (6-23) 对于舍入误差确实比一阶方程的相应方法更为敏感。

6.3-2. *一个后验界*. 现在对 $|r_n|$ 导出一个界，它至少在所有 $|\varepsilon_n|$ 有指定的极大值的情形下是(渐近地)严格的。推导 $h \to 0$ 的极限过程，必须假设 $Nh^{p+2} \leqslant \varepsilon \leqslant Kh^3$（注意按指数变化）。由定理 6.7 能断言 $r_m = O(h)$。我们还假设 $f_{yy}(x,y)$ 存在且在精确解 $y = y(x)$ 的一个邻域内连续。若 h 充分小，则有

$$f(x_m, \tilde{y}_m) - f(x_m, y_m) = g(x_m) r_m + \theta_m K_1 \varepsilon h^{-1},$$

其中 $g(x) = f_y(x, y(x))$。因此，我们便能将 (6-102) 换成

$$\alpha_k r_{n+k} + \cdots + \alpha_0 r_n = h^2\{\beta_k g_{n+k} r_{n+k} + \cdots + \beta_0 g_n r_n\} + \varepsilon_{n+k} + \theta_n K_2 h \varepsilon, \tag{6-104}$$

这里 g_m 的意义现在就不同了：$g_m = g(x_m)$，相应地，我们将 r_n 分成和 $r_n^{(1)} + r_n^{(2)}$，其中 $r_n^{(1)}$ 称为主要误差，并定义为 $\theta_n = 0$ 时(6-104)的解，而此处的 $r_n^{(2)}$ 称为次要误差，它定义为 $\varepsilon = 0$，$\theta_n \neq 0$ 时的解。应用引理 6.3，便能导出 $r_n^{(2)}$ 的一个界，$r_n^{(2)} = O(\varepsilon h^{-1})$。另一方面，主要误差一定期望为 $O(\varepsilon h^{-2})$。因此，当 $h \to 0$ 时 r_n 的性态由主要误差来决定。为此，本节余下的部分仅考虑这个主要误差。

主要误差定义是初值为零的线性非齐次差分方程的解。按照定理 5.2，它可以表示为如下形式：

$$r_n^{(1)} = \sum_{l=k}^{n} \varepsilon_l d_{n,l}, \tag{6-105}$$

其中对于 $l = k, k+1, \cdots, \{d_{n,l}\}$ 是差分方程

$$\alpha_k d_{n+k,l} + \alpha_{k-1} d_{n+k-1,l} + \cdots + \alpha_0 d_{n,l} = h^2\{\beta_k g_{n+k} d_{n+k,l} + \cdots + \beta_0 g_n d_{n,l}\} \tag{6-106}$$

取初值

$$d_{n,l} = \begin{cases} 0, & n < l, \\ \dfrac{1}{\alpha_k - h^2\beta_k g_l}, & n = l \end{cases} \tag{6-107}$$

的解.

考察离散误差 e_n (参看 §6.2-3) 所满足的差分方程，这表明，如果 $C = 0$，$d_{n,l}$ 的初值问题等价于 e_n 的初值问题，若 x_0 换成 x_{l-k+1}，并且初始条件按下面的特殊方式选取:

$$\delta_\mu(h) = 0, \quad \mu = 0, 1, \cdots, k-2,$$

$$\delta_\mu(h) = \frac{h^r}{\alpha_k - h^2\beta_k g_l}, \quad \mu = k-1.$$

因此，$d_{n,l}$ 的渐近性态在作了一个适当的变量变换后可从 (6-100) 得到． 由于 $\rho_\nu(\zeta) = \rho(\zeta)/(\zeta - \zeta_\nu)$，于是得到 $\alpha_{\nu,k-1} = \alpha_k$． 因此

$$h^{-r-1}\{\alpha_{\nu,k-1}\delta_{k-1}(h) + \cdots + \alpha_{\nu,0}\delta_0(h)\} = \frac{1}{h} + O(1)$$

与 (6-100) 便导出

$$d_{n,l} = \sum_{k=1}^{d} \frac{2}{h\rho''(\zeta_{2k})} \exp[i(n-l+k-1)\varphi_{3k}] d_{l,k}(x_n) + O(1), \tag{6-108}$$

这里函数 $d_{l,k}(x)$ 定义为初值问题

$$\begin{aligned} d''_{l,k}(x) &= \mu_{2k}^2 g(x) d_{l,k}(x), \\ d_{l,k}(x_l) &= 0, \quad d'_{l,k}(x_l) = 1 \end{aligned} \tag{6-109}$$

的解.

下一步将以(6-99)所定义的函数 $s_k(x)$ 以及初值问题

$$c_k''(x) = \mu_{2k}^2 g(x) c_k(x),$$
$$c_k(a) = 1,\ c_k'(a) = 0, k = 1,2,\cdots,d \tag{6-110}$$

的解所定义的类似函数 $c_k(x)$ 来表示函数 $d_{l,k}(x)$. 我们断言

$$d_{l,k}(x) = f_k(x_l, x), \tag{6-111}$$

其中当 $a \leqslant t \leqslant b$, $a \leqslant x \leqslant b$ 时,

$$f_k(t,x) = c_k(t)s_k(x) - s_k(t)c_k(x). \tag{6-112}$$

我们将证明,如果函数 $d_{l,k}(x)$ 由(6-111)所定义,则它们满足条件(6-109). 对每一个固定的 t 值, $f_k(t,x)$ 为微分方程(6-109)的解的线性组合,因而它自己也是一个解. (6-109)中二个初始条件的第一个由于 $f_k(x,x) = 0$ 而明显地满足. 为了证明第二个条件亦满足,我们指出

$$\frac{d}{dx} f_k(t,x)|_{x=t} = W(t),$$

其中

$$W(t) = c_k(t)s_k'(t) - s_k(t)c_k'(t).$$

我们有 $W(a) = 1$ 且由于

$$W_k'(t) = c_k(t)s_k''(t) - s_k(t)c_k''(t)$$
$$= \mu_{2k}^2 g(t)[c_k(t)s_k(t) - s_k(t)c_k(t)] = 0,$$

故当 $t \geqslant a$ 时 $W(t) = 1$. 这便建立了所要的结果. 利用(6-105), (6-108)以及(6-111), 因而便得到基本公式

$$r_n^{(1)} = \frac{1}{h} \sum_{l=k}^{n} \varepsilon_l \left\{ \sum_{k=1}^{d} \frac{2\exp[i(n-l+k-1)\varphi_{2k}]}{\rho''(\zeta_{2k})} f_k(x_l, x_n) + O(h) \right\}. \tag{6-113}$$

我们将首先应用它来获得 $|r_n^{(1)}|$ 的一个渐近界. 假设

$$|\varepsilon_l| \leqslant p(x_l)\varepsilon, \tag{6-114}$$

其中 $\varepsilon \geqslant 0$ 与 x 和 l 无关,而 $p(x)$ 为一个已知的分段连续函

数，它与 h 无关．重新安排二重和的次序，从（6-113）得到

$$|r_n^{(1)}| \leqslant \frac{2\varepsilon}{h^2} \sum_{k=1}^{d} |\rho''(\zeta_{2k})|^{-1} m_{k,n},$$

其中

$$m_{k,n} = h \sum_{l=k}^{n} \{p(x_l)|f_k(x_l, x_n)| + O(h)\}.$$

由引理 5.8（以积分近似和式）有

$$m_{k,n} = m_k(x_n) + O(h),$$

其中

$$m_k(x) = \int_a^x p(t)|f_k(t, x)|\,dt. \qquad (6\text{-}115)$$

现在得到下面的结果：

定理 6.9. 如果局部舍入误差满足（6-114），其中

$$\varepsilon = O(h^2),$$

那么

$$|r_n^{(1)}| \leqslant \frac{2\varepsilon}{h^2} \sum_{k=1}^{d} \{|\rho''(\zeta_{2k})|^{-1} m_k(x_n) + O(h)\},$$

$$(6\text{-}116)$$

其中函数 $m_k(x)$ 由（6-115）确定．

可以导出用微分方程的解来表示函数 $m_k(x)$ 的不等式，见问题 17 和 18．

6.3-3．统计理论．现在转向舍入误差的统计模型，假设局部舍入误差为独立的随机变量，且满足

$$|E(\varepsilon_l)| \leqslant \mu p(x_l), \qquad (6\text{-}117)$$

$$\operatorname{var}(\varepsilon_l) = \sigma^2 q(x_l), \qquad (6\text{-}118)$$

这里 $p(x)$ 及 $q(x)$ 在 $[a, b]$ 上为非负分段光滑函数，而 μ 和 σ^2 为仅与 h 有关的非负常数． 注意（以前各章也许已做过类似的注释）这个模型包含了 §6.3-2 中作为一种特殊情形 $\sigma^2=0$，

$\mu=\varepsilon$ 所开发的舍入误差的关键性研究．因此，下面的理论允许一个从统计到非统计方法的连续过渡．以后我们将讨论在具体情形下，μ，σ^2，$p(x)$及 $q(x)$应该怎样选取．

关于 $E(r_n^{(1)})$ 的界是容易计算的．从（6-113)，并用（194)，得到

$$E(r_n^{(1)})=\frac{1}{h}\sum_{l=k}^{n}E(\varepsilon_l)\left\{\sum_{k=1}^{d}\frac{2\exp[i(n-l+k-1)\varphi_{2k}]}{\rho''(\zeta_{2k})}\right.$$

$$\left.\times f_k(x_l,x_n)+O(h)\right\}. \qquad (6\text{-}119)$$

应用（6-118）并利用函数 $m_k(x)$的定义（6-115)，我们得到

$$|E(r_n^{(1)})|\leqslant\frac{2\mu}{h^2}\left\{\sum_{k=1}^{d}|\rho''(\zeta_{2k})|^{-1}m_k(x_n)+O(h)\right\}. \qquad (6\text{-}120)$$

为了计算方差，我们仍从（6-105）开始，得到

$$\operatorname{var}(r_n^{(1)})=\frac{\sigma^2}{h^3}v_n,\quad v_n=h^3\sum_{l=k}^{n}q_l d_{nl}^2, \qquad (6\text{-}121)$$

其中 $q_l=q(x_l)$，并由（6-108)，有

$$hd_{nl}=\sum_{k=1}^{d'}\frac{2\exp[i(n-l+k-1)\varphi_{2k}]}{\rho''(\zeta_{2k})}f_k(x_l,x_n)+O(h)$$

由于 d_{nl} 是实的，我们有

$$h^2d_{nl}^2=h^2d_{nl}d_{nl}$$

$$=\sum_{k=1}^{d}c_{kk}d_{nlkk}+\sum_{\substack{k,\lambda=1\\k\neq\lambda}}^{d}c_{k\lambda}\exp[i(n-l+k-1)\delta_{k\lambda}]d_{nlk\lambda}$$

$$+O(h),$$

其中 $k,\lambda=1,2,\cdots,d$，

$$\begin{aligned}c_{k\lambda}&=[\rho''(\zeta_{2k})\overline{\rho''(\zeta_{2\lambda})}]^{-1},\\ \delta_{k\lambda}&=\varphi_{2k}-\varphi_{2\lambda},\\ d_{nlk\lambda}&=f_k(x_l,x_n)\overline{f_\lambda(x_l,x_n)}.\end{aligned} \qquad (6\text{-}122)$$

对换求和的顺序，得到

$$v_n = \sum_{k=1}^{d} c_{kk} v_{nkk} + \sum_{\substack{k,\lambda=1 \\ k \neq \lambda}}^{d} c_{k\lambda} v_{nk\lambda} \exp[i(n - l + k - 1)\delta_{k\lambda}] + O(h), \tag{6-123}$$

其中

$$v_{nk\lambda} = h \sum_{l=k}^{n} q_l d_{nlk\lambda}.$$

如 $k = \lambda$，由引理 5.8，我们得到(用积分代替求和)

$$v_{nkk} = v_k(x_n) + O(h), \tag{6-124}$$

其中

$$v_k(x) = \int_a^x q(t) |f_k(t, x)|^2 dt. \tag{6-125}$$

当 $k \neq \lambda$，由引理 5.9 (离散型类似于 Riemann-Lebesgue 引理)，得 $v_{nk\lambda} = O(h)$. 因此，对 v_n 来说唯一起决定作用的是第一个和，所以我们可以叙述为如下的结果:

定理 6.10. 如果局部舍入误差为满足 (6-117) 及 (6-118) 的独立随机变量，那么累积舍入误差的主要分量为一个随机变量，其均值满足 (6-120)，方差由

$$\operatorname{var}(r_n^{(1)}) = \frac{4\sigma^2}{h^3} \sum_{k=1}^{d} \{ |\rho''(\zeta_{2k})|^{-2} v_k(x_n) + O(h) \} \tag{6-126}$$

给出，其中 $v_k(x)$ 由 (6-125) 确定.

若当 $a \leqslant x \leqslant b$ 时 $q(x) > 0$，还可证明对中心极限定理起保证作用的条件 (1-96) 是满足的.

这个结果说明在统计模型中标准偏差与一阶的情形 $O(h^{-1/2})$ 相比为 $O(h^{-3/2})$，因而对通常的积分是不重要的. 这个结果还表明当 $h \to 0$ 时舍入误差的性态与开始误差一样主要依赖于 $\rho(\zeta)$ 的重本性根和有关的增长参数.

可能会有这样的问题，在二阶的情形下，关于函数 $v_k(x)$ 的微分方程 (5-254) 是怎样的。我们将在假设 μ_{2k}^2 为实的情形下来答回这个问题。因而函数 $s_k(x)$ 和 $c_k(x)$ 也是实的，且 $\overline{f_k(t,x)} = f_k(t,x)$。为了简单起见，省略下标 k 并且对 x 的微分以“ ' ”来表示，便有

$$
\begin{aligned}
v'(x) &= \int_a^x q(t)[f(t,x)^2]' dt, \\
v''(x) &= \int_a^x q(t)[f(t,x)^2]'' dt.
\end{aligned} \tag{6-127}
$$

并且由 $f'(x,x) = 1$ 得到

$$
v'''(x) = q(x) + \int_a^x q(t)[f(t,x)^2]''' dt.
$$

又由于 $(f^2)' = 2ff'$，$(f^2)'' = 2(f')^2 + 2ff'' = 2f'^2 + 2\mu^2 gf^2$，便有

$$
(f^2)''' = 4\mu^2 g(f^2)' + 2\mu^2 g' f^2.
$$

从上面的关系式以及从 (6-127) 中的积分当 $x = a$ 时变为零，就得出 $v_k(x)$ 可以认为是初值问题：

$$
\begin{aligned}
&v_k'''(x) = 4\mu_{2k}^2 g(x) v_k'(x) + 2\mu_{2k}^2 g'(x) v_k(x) + 2q(x), \\
&v_k(a) = v_k'(a) = v_k''(a) = 0
\end{aligned} \tag{6-128}
$$

的解。

作为定理 6.10 的一个应用，我们来考察方程

$$
y'' = -\omega^2 y \tag{6-129}
$$

的解 (ω 为实的)，采用二种方法：

(a) $y_{n+1} - 2y_n + y_{n-1} = h^2 f_n$;

(b) $y_{n+2} - 2y_n + y_{n-2} = \frac{4}{3} h^2 (f_{n+1} + f_n + f_{n-1})$

[见 (6-9) 和 (6-18)] 假设积分从 $x = 0$ 开始，初始条件的确切形式是不重要的。

对于方法 (a)，我们有 $\rho(\zeta) = (\zeta - 1)^2$，$\sigma(\zeta) = \zeta$，$d =$

$1, \zeta_1 = \zeta_2 = 1, \mu_2^2 = 1$. 因此在(6-126)中的和便化成含有函数 $v_1(x)$ 的项，而 $v_1(x)$ 定义为

$$v_1''' = -4\omega^2 v_1' + 2q, \quad v(0) = v'(0) = v''(0) = 0$$

的解。若 $q = 1$（定点十进制运算），易证其解为

$$v_1(x) = \frac{1}{4\omega^3}(2\omega x - \sin 2\omega x). \qquad (6\text{-}130)$$

因此，由于 $\rho''(1) = 2$，对于方法(a)，有

$$\operatorname{var}(r_n^{(1)}) = \frac{\sigma^2}{4\omega^3 h^3}(2\omega x - \sin 2\omega x + O(h)).$$

舍入误差的增长对 x 而言近乎线性的［正如在积分(6-129)一样，将它化成一阶方程组并应用单步方法］。

对于方法(b)，我们得到 $\rho(\zeta) = \zeta^4 - 2\zeta^2 + 1$，$\sigma(\zeta) = \frac{4}{3}(\zeta^3 + \zeta^2 + \zeta)$，$d = 2$，$\zeta_1 = \zeta_2 = 1$，$\zeta_3 = \zeta_4 = -1$，$\mu_2^2 = 1$，$\mu_4^2 = -\frac{1}{3}$。(6-126)中的和包含二项。第一项含有如上确定的函数 $v_1(x)$；第二项含有由

$$v_2''' = \frac{4}{3}\omega^2 v_2' + 2q, \quad v(0) = v'(0) = v''(0) = 0$$

确定的函数 $v_2(x)$。当 $q = 1$ 时解为

$$v_2(x) = \frac{3^{3/2}}{4\omega^3}\left(\sinh\frac{2\omega}{\sqrt{3}}x - \frac{2\omega}{\sqrt{3}}x\right). \qquad (6\text{-}131)$$

利用 $\rho''(1) = \rho''(-1) = 8$，对于方法(b)，我们得到

$$\operatorname{var}(r_n^{(1)}) = \frac{\sigma^2}{64\omega^3 h^3}\left[\left(2 - \frac{2}{\sqrt{3}}\right)\omega x - \sin 2\omega x + \sinh\frac{2\omega x}{\sqrt{3}} + O(h)\right].$$

对于相同的 σ 值，开始时方差的增长方法(b)较方法(a)

略慢，稍后方法(b)由于 sinh 函数的出现，误差成指数地增长。这二条曲线于

$$\frac{1}{16}\left[\left(2-\frac{2}{\sqrt{3}}\right)\omega x-\sin 2\omega x+\sinh\frac{2\omega x}{\sqrt{3}}\right]$$
$$=2\omega x-\sin 2\omega x$$

处相交，即在区间 $5.0\leqslant \omega x\leqslant 5.1$ 的某个地方。

在上述的例题中，以及进一步对于任何具有形式 $y''=f(x,y)$ 的线性微分方程，函数 $g(x)=f_y(x,y(x))$ 与所考虑的特解 $y(x)$ 无关，因此也与 $y(a)$，$y'(a)$ 无关。一个有力的说明是数学不稳定问题 $y''=y$，$y(0)=1$，$y'(0)=-1$，其精确解 $y(x)=e^{-x}$ 按指数下降。但是以方法(a)来积分这个方程，在 $q(x)=1$ 的假设下，我们得到

$$\operatorname{var}(r_n^{(1)})=\frac{\sigma^2}{4h^3}(\sinh 2x-2x+O(h)).$$

假设 $q(x)=e^{-2x}$（对浮点运算来说它是合理的），我们有

$$\operatorname{var}(r_n^{(1)})=\frac{\sigma^2}{16h^3}[e^{2x}+(3+4x)e^{-2x}-4+O(h)],$$

指数增长的分量仍然出现。

6.3-4. **增长参数的某些性质.** 如果 ζ 为 $\rho(\zeta)$ 的一个重本性根。相应的增长参数定义为

$$\mu^2=\frac{2\sigma(\zeta)}{\zeta^2\rho''(\zeta)}.$$

我们将以多项式 $r(z)$ 及 $s(z)$ 来表示增长参数并导出它的一些性质，特别是关于最佳方法的性质。

从(5-112)通过微分，我们得到

$$\begin{aligned}\rho''(\zeta)=k(k-1)(\zeta+1)^{k-2}r(z)\\+(4k-2)(\zeta+1)^{k-3}r'(z)\\+4(\zeta+1)^{k-4}r''(z),\end{aligned}\qquad(6\text{-}132)$$

其中 $z=(\zeta-1)/(\zeta+1)$. 若根 ζ 异于 -1, 则 $r(z)=r'(z)=0$,便得到

$$\mu^2=\frac{8s(z)}{(1-z^2)^2r''(z)}. \tag{6-133}$$

如果 $\zeta=-1$ 为一个重根,多项式 $r(z)$ 具有形式

$$r(z)=a_2z^2+\cdots+a_{k-2}z^{k-2},$$

这里我们可以假设 $a_2>0$, $a_{k-2}>0$. 如果 $s(z)$ 由 (6-62) 确定,那么在 (6-132) 中令 $\zeta\to-1$, 便得

$$\mu^2=4b_k/a_{k-2}. \tag{6-134}$$

如果 $r(z)$ 的所有根均在虚轴上,那么 $r(z)$ 为偶的或为奇的函数随 k 为偶数或奇数而定。 如果多项式 $r''(z)$ 是由 (6-62) 确定,则 $s(z)$ 与 $r(z)$ 情形相同. 于是由 (6-133) 得到数 μ^2 均为实的,并且同样关于 (6-134) 来说一般也是对的.

如果 k 是偶数,方法是最佳的,并且如果 $\zeta=-1$ 为一个重根, 那么由 §6.1-8 中给出的系数 b_k 的显式表达式中可以导出关于对应的增长参数的一个有趣的不等式.利用(6-64),我们得到

$$\mu^2=\frac{4(d_4a_{k-2}+d_6a_{k-4}+\cdots+d_ka_2)}{a_{k-2}}.$$

由于 $d_{2\nu}<0(\nu=1,2,\cdots)$, $a_\nu\geqslant 0(\nu=2,\cdots,k-2)$, 再利用 $d_4=-\dfrac{1}{60}$, 便有

$$\mu^2\leqslant-\frac{1}{15}. \tag{6-135}$$

还可得到进一步的不等式

$$\mu^2\leqslant 4d_k\frac{a_2}{a_{k-2}}. \tag{6-136}$$

如果将这个不等式与关于误差常数C 的不等式 (6-68) 一起

来考虑，便看出对于 $k \geq 4$，μ^2 这个量与 C 不能同时达到它们的上界 (6-135) 和 (6-69)。

将定理 5.16 推广到二阶的情形看来是不可能的。

6.3-5. *局部舍入误差.* 局部舍入误差记为 ε_{n+k}，它满足关系式

$$\alpha_k \tilde{y}_{n+k} + \cdots + \alpha_0 \tilde{y}_n = h^2\{\beta_k f(x_{n+k}, \tilde{y}_{n+k}) + \cdots + \beta_0 f(x_n, \tilde{y}_n)\} + \varepsilon_{n+k},$$

其中量 $\tilde{y}_m (m = 0, 1, \cdots)$ 表示 y_m 的真正计算值. 对 ε_{n+k} 的任何一种研究，不论是统计的还是非统计的，都必须从这个定义出发. 我们不准备对实际中可能有的多种计算方法进行研究，然而确实希望强调前面理论中的一个重要结果. 如果采用以 u 为基本单位的定点运算，而且如果所有的量都以同样的精确度进行计算，那么 ε_{n+k} 由于在方括号中的表达式与 h^2 的最后相乘是与 u 同阶的（这种考虑忽略了固有误差和由迭代而产生的误差的这些因素只能使事情变坏）. 由 (6-126) 得到累积误差的标准偏差具有 $uh^{-3/2}$ 阶. 正如已经指出的那样，这比我们在积分一阶方程时(或一阶方程组)要差一个 h^{-1} 的因子. 另一方面，如果使用部分双倍位精确度，那么局部舍入误差具有 h^2u 阶，因为此时乘法 $h^2\{\}$ 是精确地完成的[1]. 由迭代可能产生一个附加的误差，但通过适当的预防这个误差可控制在 h^2u 以内. 由此得出，局部误差的标准偏差这时便具有 h^2u 阶，而累积误差的阶为 $uh^{1/2}$，它就是我们用部分双倍位精度积分一阶方程时得到的. 同样的考虑可应用于浮点运算. 因此看来部分双倍位精度用于二阶方程要比一阶方程相对来说更为有效.

另外，现在要讨论增加形如 (6-23) 的某些方法的数值精

---

1) 我们假设 $h^2 > u$.

确度的更为有效的方法.

6.4. 差分方程的求和形式

这一节讨论用一个新的方法计算由 (6-23) 确定的值 y_n. 如果没有舍入误差,又如果开始值是完全相同的,那么由这个新算法所产生的值 y_n 与由通常的方式所获得的是一样的.然而,从舍入误差的传播来考虑,新的算法的特性是非常不同的,至少使用的单倍位精确度运算是这样.

6.4-1. 算法陈述. 对于 $n=0,1,2,\cdots,N$, 将(6-23)写出来并记所得到的方程的和为 S_N. 左边相加,得到

$$S_N=\alpha_k(y_k+y_{k+1}+\cdots+y_{N+k})+\alpha_{k-1}(y_{k-1}+y_k+\cdots+y_{n+k-1})+\cdots+\alpha_0(y_0+y_1+\cdots+y_N),$$

或经过重排,

$$\begin{aligned}S_N=&\alpha_0y_0+(\alpha_1+\alpha_0)y_1+\cdots+(\alpha_{k-1}+\alpha_{k-2}+\cdots\\&+\alpha_0)y_{k-1}+(\alpha_k+\alpha_{k-1}+\cdots+\alpha_0)(y_k+y_{k+1}\\&+\cdots+y_N)+(\alpha_k+\alpha_{k-1}+\cdots+\alpha_1)y_{N+1}\\&+(\alpha_k+\cdots+\alpha_2)y_{N+2}+\cdots+\alpha_ky_{N+k}.\end{aligned}\tag{6-137}$$

假设方法为相容的,我们有 $\rho(1)=0$, 因而

$$\alpha_k+\alpha_{k-1}+\cdots+\alpha_0=0.\tag{6-138}$$

同时由

$$\rho_1(\zeta)=\frac{\rho(\zeta)}{\zeta-1}=\alpha'_{k-1}\zeta^{k-1}+\alpha'_{k-2}\zeta^{k-2}+\cdots+\alpha'_0\tag{6-139}$$

来定义多项式 $\rho_1(\zeta)$ 及系数 $\alpha'_{k-1},\alpha'_{k-2},\cdots,\alpha'_0$. 通过在(6-139)上乘以 $(\zeta-1)$, 比较系数,并利用 (6-138), 便得到,

$$\alpha'_{k-1}=\alpha_k=-\alpha_{k-1}-\alpha_{k-2}-\cdots-\alpha_0,$$

$$\alpha'_{k-2}=\alpha_k+\alpha_{k-1}=-\alpha_{k-2}-\cdots-\alpha_0,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\alpha_0' = \alpha_k + \alpha_{k-1} + \cdots + \alpha_1 = -\alpha_0.$$

因此(6-137)便可写成

$$S_N = \alpha_{k-1}' y_{N+k} + \alpha_{k-2}' y_{N+k-1} + \cdots + \alpha_0' y_{N+1} \\ - \alpha_{k-1}' y_{k-1} - \alpha_{k-2}' y_{k-2} - \cdots - \alpha_0' y_0. \tag{6-140}$$

将(6-23)的右边项相加,便得到

$$S_N = h^2\{\beta_k(f_k + f_{k+1} + \cdots + f_{N+k}) \\ + \beta_{k-1}(f_{k-1} + f_k + \cdots + f_{N+k-1}) \\ + \cdots + \beta_0(f_0 + f_1 + \cdots + f_N)\}. \tag{6-141}$$

由于对于一个稳定的方法来说, $\beta_k + \beta_{k-1} + \cdots + \beta_0 \neq 0$,所以,象上面所出现的简化是不成立的.然而,为了缩写起见,可以令

$$\frac{h}{\alpha}\sum_{\mu=0}^{n} f_\mu = F_n - H,\quad n = 0,1,2,\cdots, \tag{6-142}$$

其中 $\alpha \neq 0$ 为一个常量参数,它将起到一个比例因子的作用.按照这样的方式选择常数 H,使恒等式

$$\alpha_{k-1}' y_{N+k} + \alpha_{k-2}' y_{N+k-1} + \cdots + \alpha_0' y_{N+1} \\ = \alpha h\{\beta_k F_{N+k} + \beta_{k-1} F_{N+k-1} + \cdots + \beta_0 F_N\} \tag{6-143}$$

成立.对(6-140)和(6-141)进行比较,可以看出这个关系式导出

$$\alpha_{k-1}' y_{k-1} + \cdots + \alpha_0' y_0 \\ = \alpha h\{\beta_k F_{k-1} + \beta_{k-1} F_{k-2} + \cdots + \beta_1 F_0 + \beta_0 H\} \\ = h^2\{\beta_k f_{k-1} + (\beta_k + \beta_{k-1}) f_{k-2} + \cdots + (\beta_k + \beta_{k-1} \\ + \cdots + \beta_1) f_0\} + \alpha h(\beta_k + \beta_{k-1} \\ + \cdots + \beta_0) H. \tag{6-144}$$

由

$$\sigma_1(\zeta) = \frac{\sigma(\zeta) - \sigma(1)}{\zeta - 1} = \beta_{k-1}' \zeta^{k-1} + \beta_{k-2}' \zeta^{k-2} + \cdots + \beta_0'$$

来定义多项式 $\sigma_1(\zeta)$ 与系数 $\beta'_{k-1},\beta'_{k-2},\cdots,\beta'_0$，我们得到

$$\beta'_{k-1}=\beta_k,$$
$$\beta'_{k-2}=\beta_k+\beta_{k-1},$$
$$\cdots\cdots\cdots\cdots$$
$$\beta'_0=\beta_k+\beta_{k-1}+\cdots+\beta_1.$$

我们可将 (6-144) 写成形式

$$\alpha h\,\sigma(1)H=\alpha'_{k-1}y_{k-1}+\cdots+\alpha'_0y_0 - h^2\{\beta'_{k-1}f_{k-1}+\cdots+\beta'_0f_0\}, \qquad (6-145)$$

此处在右边的表达式为开始值的已知函数，并且由于

$$\sigma(1)\neq 0$$

可以毫无困难地解出 H.

方程

$$\alpha'_{k-1}y_{n+k}+\alpha'_{k-2}y_{n+k-1}+\cdots+\alpha'_0y_{n+1} = \alpha h\{\beta_kF_{n+k}+\cdots+\beta_0F_n\}, \qquad (6-146a)$$

$$F_{-1}=H,\ F_{n+k}-F_{n+k-1}=\frac{h}{\alpha}f_{n+k} \qquad (6-146b)$$

定义了差分方程 (6-23) 的求和形式. 如果 H 由 (6-145) 确定，根据推导，它们为 (6-23) 的每一个解所满足. 反之，(6-146) 的每一个解满足 (6-23)，这点可通过形成 (6-146a) 的一阶向后差分并利用 (6-146b) 而看出. 因而解 (6-23) 的过程等价于解 (6-146) 的过程，并且如果开始值相同，二种方法就得到恒等的数学上的近似值 y_n.

因而这两种方法的离散误差是相同的. 特别是，由 (6-23) 所描述的方法是收敛的，当且仅当由 (6-146) 所描述的方法是收敛的，并且在两种情形下离散误差的渐近性态均由定理 6.8 所描述. 然而，从计算的观点出发，这二种方法是显然不同的. 在 §6.4-2 及 6.4-3 中将讨论这些差别影响舍入

误差传播的问题。

对于 Störmer 及 Cowell 方法来说，公式（6-146a）及（6-145）有一个特别简单的状态。在这两种情形，$\rho(\zeta)=\zeta^k-2\zeta^{k-1}+\zeta^{k-2}$，因而 $\rho_1(\zeta)=\zeta^{k-1}-\zeta^{k-2}$，$\alpha'_{k-1}=1$，$\alpha'_{k-2}=-1$，$\alpha'_{k-3}=\cdots=\alpha'_0=0$。

对于 Störmer 方法，我们有

$$\sigma(\zeta)=\sigma_0\zeta^{k-1}+\sigma_1\zeta^{k-2}(\zeta-1)+\cdots+\sigma_{k-1}(\zeta-1)^{k-1},$$

其中系数 σ_i 由（6.7）定义。我们有 $\sigma(1)=\sigma_0=1$ 以及

$$\frac{\sigma(\zeta)-\sigma(1)}{\zeta-1}=\sigma_1\zeta^{k-2}+\sigma_2\zeta^{k-3}(\zeta-1)+\cdots+\sigma_{k-1}(\zeta-1)^{k-2}.$$

为方便起见，将 $n+k-1$ 换成 m，于是可将"求和"形式的 Störmer 方法归结为公式

$$\alpha hH=\nabla y_{k-1}-h^2\{\sigma_1 f_{k-2}+\sigma_2\nabla f_{k-2}+\cdots+\sigma_{k-1}\nabla^{k-2}f_{k-2}\},\tag{6-147a}$$

$$F_{-1}=H,\ \nabla F_m=\frac{h}{\alpha}f_m,\tag{6-147b}$$

$$\nabla y_{m+1}=\alpha h\{\sigma_0 F_m+\sigma_1\nabla F_m+\cdots+\sigma_{k-1}\nabla^{k-1}F_m\}.\tag{6-147c}$$

它的实际应用如下：我们从开始值计算 $f_0,f_1,\cdots,f_{k-1}$，接着由（147a）确定出 H。然后从（6-147b）计算 $F_0,F_1,\cdots F_{k-1}$，于是便形成了差分。y_k 就由（6-147c）来计算，这就使我们得到 $f_k=f(x_k,y_k)$ 及 F_k。这样可得到新的差分 $\nabla^q F_k$ 并由（6-147c）确定出 y_{k+1}。从这里向前计算，循环进行。建立在显式公式上的其它方法，其求和形式的计算过程也是类似的。

对于 Cowell 方法，

$$\sigma(\zeta)=\sigma_0^*\zeta^k+\sigma_1^*\zeta^{k-1}(\zeta-1)+\cdots+\sigma_k^*(\zeta-1)^k,$$

其中系数 σ_i^* 由 (6-12) 给出. 因此

$$\frac{\sigma(\zeta)-\sigma(1)}{\zeta-1}=\sigma_1^*\zeta^{k-1}+\sigma_2^*\zeta^{k-2}(\zeta-1)+\cdots$$
$$+\sigma_k^*(\zeta-1)^{k-1}.$$

将 $n+k$ 记成 m, 与公式 (6-147) 的相应部分为

$$\alpha hH=\nabla y_{k-1}-h^2\{\sigma_1^*f_{k-1}+\sigma_2^*\nabla f_{k-1}+\cdots$$
$$+\sigma_k^*\nabla^{k-1}f_{k-1}\}, \qquad (6\text{-}148a)$$

$$F_{-1}=H,\ \nabla F_m=\frac{h}{\alpha}f_m, \qquad (6\text{-}148b)$$

$$\nabla y_m=\alpha h\{\sigma_0^*F_m+\sigma_1^*\nabla F_m+\cdots+\sigma_k^*\nabla^k F_m\}. \qquad (6\text{-}148c)$$

这些公式的应用类似于 (6-147), 重要的差别是由于 F_m 也依赖于 y_m, 所以方程 (6-148c) 为未知量 y_m 的一个隐式方程. (若方法为隐式的, 这个性质通过求和过程并不能去掉) 因此只需使用 F_{m-1} 的向后差分公式便可以预估一个 y_m 值. 公式 (6-147c) (m 向后移一个单位) 本身就可以自然地用来为这一目的服务. 从 (6-148b) 便可得到 F_m 的一个试探值, 然后由 (6-148c) 计算 y_m 的校正值. 如果需要, 可重复这个过程,它类似于 Adams Moulton 方法(见 §5.1-2). 在(6-147) 和 (6-148) 中常数 α 可以根据计算的方便任意选取.

6.4-2. 舍入误差的界. 从 (6-146) 真正计算得到的数值 $\tilde{y}_n$ 和 $\tilde{F}_n$ 满足关系式

$$\alpha_{k-1}'\tilde{y}_{n+k}+\cdots+\alpha_0'\tilde{y}_{n+1}=\alpha h\{\beta_k\tilde{F}_{n+k}+\cdots$$
$$+\beta_0\tilde{F}_n\}+\varepsilon_{n+k},$$

$$\tilde{F}_{n+k}-\tilde{F}_{n+k-1}=\frac{h}{\alpha}f(x_{n+k},\tilde{y}_{n+k})+\eta_{n+k}, \qquad (6\text{-}149)$$

其中 ε_{n+k} 和 η_{n+k} 分别为计算 y_{n+k} 和 F_{n+k} 时的局部舍入误差. 如果使用定点运算, ε_{n+k} 和 η_{n+k} 与基本单位 u 同一个数

量级．令 $r_n = \tilde{y}_n - y_n$, $R_n = \tilde{F}_n - F_n$, 从 (6-149) 中减去相应的关系式 (6-146)，便得到

$$\alpha'_{k-1} r_{n+k} + \cdots + \alpha'_0 r_{n+1} = \alpha h \{\beta_k R_{n+k} + \cdots + \beta_0 R_n\} + \varepsilon_{n+k}, \qquad (6-150)$$

$$R_{n+k} - R_{n+k-1} = \frac{h}{\alpha} g_{n+k} r_{n+k} + \eta_{n+k}, \qquad (6-151)$$

其中

$$g_m = \begin{cases} \dfrac{f(x_m, \tilde{y}_m) - f(x_m, y_m)}{\tilde{y}_m - y_m}, & \tilde{y}_m \neq y_m, \\ 0, & \tilde{y}_m = y_m. \end{cases} \quad m = 0, 1, 2, \cdots.$$

利用 Lipschitz 条件，$|g_m| \leqslant L$, $m = 0, 1, 2, \cdots$.

现在在

$$|\varepsilon_n| \leqslant \varepsilon, \ |\eta_n| \leqslant \eta, \ n = 1, 2, \cdots \qquad (6-152)$$

的假设下导出 $|r_n|$ 的一个界．可以证明，如果 $d = 1$，亦即如果 $\zeta = 1$ 为 $\rho(\zeta)$ 在单位圆上的唯一重根，那么这个界具有 $h^{-1} \max(\varepsilon, \eta)$ 的阶．这说明它比 (6-103) 所表示的算法的标准形式改进了一个 h 因子．

我们将需要引理 6.2 的下面的改进：

引理 6.4. 如果多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 定义一个相容且稳定的方法，而且如果 $\zeta = 1$ 为 $\rho(\zeta)$ 满足 $|\zeta| = 1$ 的唯一重根，那么存在一个常数 $\gamma > 0$，使得由

$$\frac{1}{\alpha_k + \alpha_{k-1}\zeta + \cdots + \alpha_0 \zeta^k} = \gamma_0 + \gamma_1 \zeta + \gamma_2 \zeta^2 + \cdots$$

确定的系数 $\gamma_n (n = 0, 1, 2, \cdots)$ 都具有形式

$$\gamma_n = \frac{n}{\sigma(1)} + \gamma'_n, \qquad (6-153)$$

其中 $|\gamma'_n| \leqslant \gamma$.

证．令 $\alpha_k + \alpha_{k-1}\zeta + \cdots + \alpha_0 \zeta^k = \hat{\rho}(\zeta)$．从 $\rho(\zeta)$ 的

相容性很快得到 $\hat{\rho}(1)=\hat{\rho}'(1)=0$，$\hat{\rho}''(1)=2\sigma(1)$. 由于 $\sigma(1)\neq 0$，$1/\hat{\rho}(\zeta)$ 以部分分式的展开式其形式为

$$\frac{1}{\hat{\rho}(\zeta)}=\frac{1}{(\zeta-1)^2\sigma(1)}+f(\zeta), \qquad (6\text{-}154)$$

其中 $f(\zeta)$ 为一个有理函数，它的极点 ζ_i 为单重而且满足

$$|\zeta_i|\geqslant 1.$$

由于在引理 5.6 的证明中所采用的理由，因而 $f(\zeta)$ 在 $\zeta=0$ 处的 Taylor 展式的系数是有界的，(6-154) 的右边第一项的 Taylor 展式中 ζ^n 的系数为 $(n+1)/\sigma(1)$. 关系式 (6-153) 现在便显然了.

例. 对于 Störmer 和 Cowell 方法，

$$\rho(\zeta)=\zeta^k-2\zeta^{k-1}+\zeta^{k-2},\ \sigma(1)=1.$$

因此(6-154)中的函数 $f(\zeta)$ 为零，取 $\gamma=1$，则 (6-153) 成立.

现在叙述本节的主要结果:

定理 6.11. 如果 $d=1$ 并且局部舍入误差满足(6-152)，那么用算法 (6-146) 对 $y''=f(x,y)$ 求数值解的累积舍入误差当 $h^2\leqslant L^{-1}|\alpha_k\beta_k^{-1}|$时满足

$$r_q\leqslant\left\{\alpha\beta\Gamma^*(x_q-a^*)^2\frac{\eta}{2h}+(\Gamma^*+2\gamma^*)(x_q-a)\frac{\varepsilon}{h}\right\}$$
$$\times\exp[(x_q-a^*)^2\Gamma^*LB],\ a\leqslant x_q\leqslant b, \quad (6\text{-}155)$$

其中 B 和 a^* 像在定理 6.7 中一样确定，其中

$$\Gamma^*=\frac{\sigma(1)^{-1}}{1-h^2L|\alpha_k^{-1}\beta_k|},\ \gamma^*=\frac{\gamma}{1-h^2L|\alpha_k^{-1}\beta_k|}.$$

证. 应用算子∇于关系式 (6-150)，并利用 (6-151) 来表示∇R_{n+k}，我们得到

$$\alpha_k r_{m+k}+\alpha_{k-1}r_{m+k-1}+\cdots+\alpha_0 r_m=h^2\{\beta_k g_{m+k}r_{m+k}$$
$$+\cdots+\beta_0 g_m r_m\}+\alpha h\{\beta_k\eta_{m+k}+\cdots+\beta_0\eta_m\}$$
$$+\varepsilon_{m+k}-\varepsilon_{m+k-1};\qquad m=0,1,2,\cdots. \qquad (6\text{-}156)$$

将引理 6.3 的证明中所采用的技巧用于这个关系式，把(6-156)式乘以 r_{n-k-m}，对于 $m=0,1,\cdots,n-k$，并将其相加．由于 $r_0=r_1=\cdots=r_{k-1}=0$，便得到

$$
\begin{aligned}
r_n = h^2 & \Big\{\beta_k\gamma_0 g_n r_n + (\beta_k\gamma_1 + \beta_{k-1}\gamma_0)g_{n-1}r_{n-1} + \cdots \\
& + (\beta_k\gamma_{k-1} + \beta_{k-1}\gamma_{k-2} + \cdots + \beta_1\gamma_0)g_{n-k+1}r_{n-k+1} \\
& + \sum_{m=k}^{n-k}(\beta_k\gamma_m + \beta_{k-1}\gamma_{m-1} + \cdots + \beta_0\gamma_{m-k})g_{n-m}r_{n-m}\Big\} \\
& + \alpha h\Big\{\beta_k\gamma_0\eta_n + (\beta_k\gamma_1 + \beta_{k-1}\gamma_0)\eta_{n-1} + \cdots \\
& + (\beta_k\gamma_{k-1} + \cdots + \beta_1\gamma_0)\eta_{n-k+1} \\
& + \sum_{m=k}^{n-k}(\beta_k\gamma_m + \cdots + \beta_0\gamma_{m-k})\eta_{n-m}\Big\} \\
& + \gamma_0\varepsilon_n + \sum_{m=1}^{n-k}(\gamma_m - \gamma_{m-1})\varepsilon_{n-m}.
\end{aligned}
$$

利用(6-151)，(6-152)和(6-153)，右边的主要三项的界分别为

$$
h^2|\beta_k\alpha_k^{-1}|L|r_n| + h^2LB(n|\sigma(1)|^{-1}+\gamma)\sum_{m=1}^{n-k}|r_{n-m}|,
$$

$$
\alpha hB\left(\frac{1}{2}n^2|\sigma(1)|^{-1}+n\gamma\right)\eta,\quad n(2\gamma+|\sigma(1)|^{-1})\varepsilon.
$$

如果 $a\leqslant x_n\leqslant x_q$，那么有 $n\leqslant(x_q-a)h^{-1}$．于是用定理中的记号，便得到

$$
|r_n|\leqslant h(x_n-a^*)\Gamma^*LB\sum_{m=0}^{n-1}|r_m|+K,
$$

其中 K 表示(6-155)中花括号内的表达式．由此应用归纳法并应用 Bernoulli 不等式就象在引理 6.3 的证明时一样便得(6-155)．

6.4-3. **舍入误差的统计理论.** 如果 $\varepsilon < Ch^2, \eta < Dh^2$，从定理 6.11 可以得出结论 $|r_n| < Kh$。因此，对某个常数 K_1，

$$R_{n+k} - R_{n+k-1} = \frac{h}{\alpha} g_{n+k} r_{n+k} + \eta_{n+k} + \theta_n K_1 h^3, \tag{6-157}$$

这里 g_m 的定义已变成 $g_m = f_y(x_m, y(x_m))$. 微分 (6-150) 并利用 (6-157)，得到

$$\begin{aligned}\alpha_k r_{n+k} + \alpha_{k-1} r_{n+k-1} + \cdots + \alpha_0 r_n &= h^2\{\beta_k g_{n+k} r_{n+k} \\ &+ \cdots + \beta_0 g_n r_n\} + \alpha h\{\beta_k \eta_{n+k} + \cdots + \beta_0 \eta_n\} \\ &+ \varepsilon_{n+k} - \varepsilon_{n+k-1} + \theta_n K_2 h^4,\end{aligned} \tag{6-158}$$

其中 K_2 为另一个适当的常数. 相应地，我们可以将 r_n 写成主要舍入误差与一个次要舍入误差之和

$$r_n = r_n^{(1)} + r_n^{(2)},$$

其中 $r_n^{(1)}$ 被定义为 (6-158) 对应于 $\theta_n = 0$ 的解，而 $r_n^{(2)}$ 为对应于 $\varepsilon_n = \eta_n = 0$ 的解. 由引理 6.4 得到次要误差 $r_n^{(2)}$ 的界为常数乘以 h^2，于是它就具有一个单个的局部舍入误差数量级. 另一方面，主要舍入误差一定期望为 h 阶的，因此，我们将只考虑主要误差.

由于 (6-158) (对 $\theta_n = 0$)，除了一个形式更为复杂的非齐次项外，恒等于 (6-102)，并由于对这二种情形均有

$$r_0 = r_1 = \cdots = r_{k-1} = 0,$$

故有

$$r_n^{(1)} = \sum_{l=k}^{n} [\alpha h(\beta_k \eta_l + \cdots + \beta_0 \eta_{l-k}) + \varepsilon_l - \varepsilon_{l-1}] d_{n,l},$$

其中 $d_{n,l}$ 由 (6-106) 定义. 重排后，得到

$$\begin{aligned}r_n^{(1)} = &\sum_{l=k}^{n-1} (d_{n,l} - d_{n,l+1})\varepsilon_l + d_{n,n}\varepsilon_n \\ &+ \alpha h \sum_{l=k}^{n} (\beta_0 d_{n,l+k} + \beta_1 d_{n,l+k-1} + \cdots + \beta_k d_{n,l})\eta_l.\end{aligned}$$

我们仍假设仅有一个必有的本性重根在 $\zeta=1$ 处. 因此 $d=1$, 并由 (6-108) 以及其后的式子, 我们得到

$$d_{n,l}=\frac{2}{h\rho''(1)}f(x_l,x_n)+O(1), \qquad (6-159)$$

其中

$$f(t,x)=c(t)s(x)-s(t)c(x),$$

$c(x)$和 $s(x)$分别定义[1]为初值问题

$$\begin{aligned} c''&=g(x)c,\ c(a)=1,\ c'(a)=0,\\ s''&=g(x)s,\ s(a)=0,\ s'(a)=1, \end{aligned}\quad g(x)=f_y(x,y(x))$$

的解.

如果在 $d=1$ 的条件外加上该方法的 $m=2$ (即 $\zeta=1$ 为仅有的本性根, 例如对于 Störmer 以及 Cowell 方法就是那种情形). 通过一个冗长的计算可以证明, 在(6-159)中以 $O(1)$ 表示的项具有形式 $k(x_l,x_n)+O(h)$, 这个计算过程我们不在这里复述, 这里 k 为它的二个自变量的连续可微函数. 于是有 $\left(利用 \frac{1}{2}\rho''(1)=\sigma(1)=\beta_0+\cdots+\beta_k\right)$:

$$d_{n,l}-d_{n,l+1}=-\frac{2}{\rho''(1)}\frac{\partial f}{\partial t}(x_l,x_n)+O(h),$$

$$\alpha h(\beta_0 d_{n,l+k}+\cdots+\beta_k d_{n,l})=\alpha f(x_l,x_n)+O(h).$$

因此我们得到

$$r_n^{(1)}=-\frac{2}{\rho''(1)}\sum_{l=k}^{n}\left(\frac{\partial f}{\partial t}(x_l,x_n)+O(h)\right)\varepsilon_l$$

$$+\alpha\sum_{l=k}^{n}(f(x_l,x_n)+O(h))\eta_l. \qquad (6-160)$$

假设 ε_n 和 η_n 为相互独立的随机变量, 其方差为

$$\operatorname{var}(\varepsilon_n)=q(x_n)\sigma^2,\quad \operatorname{var}(\eta_n)=Q(x_n)\tau^2, \quad (6-161)$$

1) 这些函数在 §6.3 中记为 f_1, c_1 及 s_1.

其中 $q(x)$ 与 $Q(x)$ 为固定的分段连续可微函数(不依赖于 h)，而 σ 和 τ 为与 x 无关的量[但是在某种意义上，它依赖于 h 与假设 $\varepsilon_n = O(h^2)$，$\eta_n = O(h^2)$ 相容]. 于是 $r_n^{(1)}$ 的方差可用通常为形式计算给出:

$$\operatorname{var}(r_n^{(1)}) = \frac{1}{h}\left(\frac{2}{\rho''(1)}\right)^2 (t_n + O(h))\sigma^2 + \frac{\alpha^2}{h}(v_n + O(h))\tau^2,$$

其中

$$t_n = h\sum_{l=k}^{n}\left[\frac{\partial f}{\partial t}(x_l, x_n)\right]^2 q(x_l),$$

$$v_n = h\sum_{l=k}^{n}[f(x_l, x_n)]^2 Q(x_l).$$

用积分来近似和式，得到

$$t_n = t(x_n) + O(h),\ v_n = v(x_n) + O(h),$$

其中

$$t(x) = \int_a^x \left[\frac{\partial f}{\partial t}(t, x)\right]^2 q(t)\,dt \qquad (6\text{-}162)$$

以及

$$v(x) = \int_a^x [f(t, x)]^2 Q(t)\,dt. \qquad (6\text{-}163)$$

于是我们便得到下面的定理.

定理 6.12. 如果 $\zeta = 1$ 是所定义线性多步方法 (6-23) 的多项式 $\rho(\zeta)$ 的唯一本性根，而在求和形式 (6-146) 算法中的局部离散误差是满足 (6-161) 的相互独立的随机变量，那么累积舍入误差的主要分量满足

$$\operatorname{var}(r^{(1)}) = \frac{1}{h}\left(\frac{2\sigma}{\rho''(1)}\right)^2 \{t(x_n) + O(h)\} + \left(\frac{\alpha\tau}{h}\right)^2 \{v(x_n) + O(h)\}, \qquad (6\text{-}164)$$

其中函数 $t(x)$ 及 $v(x)$ 由 (6-162) 及 (6-163) 确定.

令 $k=1$ 并将 (6-128) 中的 q 换为 Q，便得到关于 $v(x)$ 的微分方程. 我们可用类似的方式得到关于 $t(x)$ 的微分方程，如果 $q(x)$ 为充分可微的话，

$$t''' - 4gt' - 2g't = q'' - 2gq, \qquad (6\text{-}165)$$

$$t(a) = 0,\ t'(a) = q(a),\ t''(a) = q'(a).$$

6.4-4. 一个数值例子. 我们已对典型问题 $y'' = -y$, $y(0) = 1$, $y'(0) = 0$ 以最简单的 Störmer 方法 [在 (6-5) 中取 $q = 0$]，并用传统格式与求和格式进行求积. 传统格式为

$$y_{n+1} - 2y_n + y_n = -h^2 y_n. \qquad (6\text{-}166)$$

由初始条件 $y_0 = 1$，并由于所要求的解关于 $x = 0$ 为对称的 (这个常识无需知道显式解，通常直接从微分方程就可以得到)，$y_1 = y_{-1}$，因而与 (6-166) 一起，便得到

$$y_1 = y_0\left(1 - \frac{1}{2}h^2\right).$$

求和形式的方法(对 $\alpha = 1$) 等价于一对方程

$$y_{n+1} - y_n = hF_n, \qquad (6\text{-}167\text{a})$$

$$F_{n+1} - F_n = -hy_{n+1}. \qquad (6\text{-}167\text{b})$$

利用上述的 y_1 值，由 (6-167a) 便得 $F_0 = \frac{1}{2}hy_0$，这对于开始计算是足够的.

数值计算是在 IBM709 计算机上以浮点运算来完成的[1]，使用对称舍入(没有用 FORTRAN 系统). 对传统的与求和的格式，计算了 500 个数值解 $\{\tilde{y}_{n,q}\}$，对应于以下的开始值:

$$\tilde{y}_{0,q} = 1 + q\Delta, \quad q = 1, 2, \cdots, 500$$

(Δ 为一个小常数)使用的步长为 $h = 2^{-7}$. 精确数值解 $\{y_{n,q}\}$

---

1) 作者对资助这些计算的 Palo Alto, California的 Lockheed 航空公司表示感谢，并且感谢将它们计算出来的 M. Dale 先生.

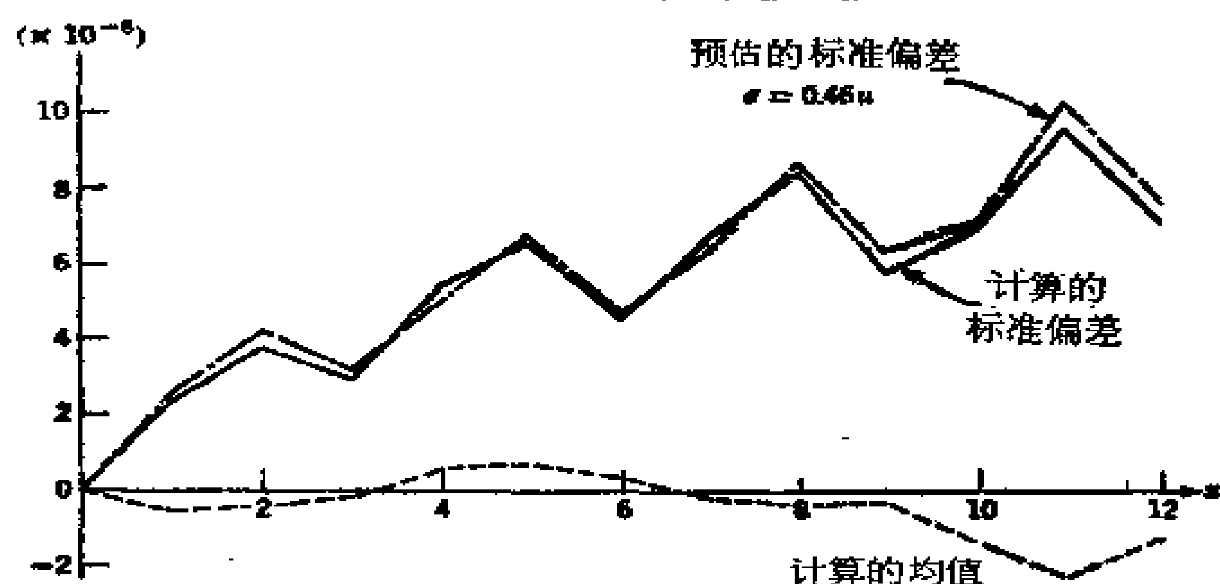


图 6.1　传统形式

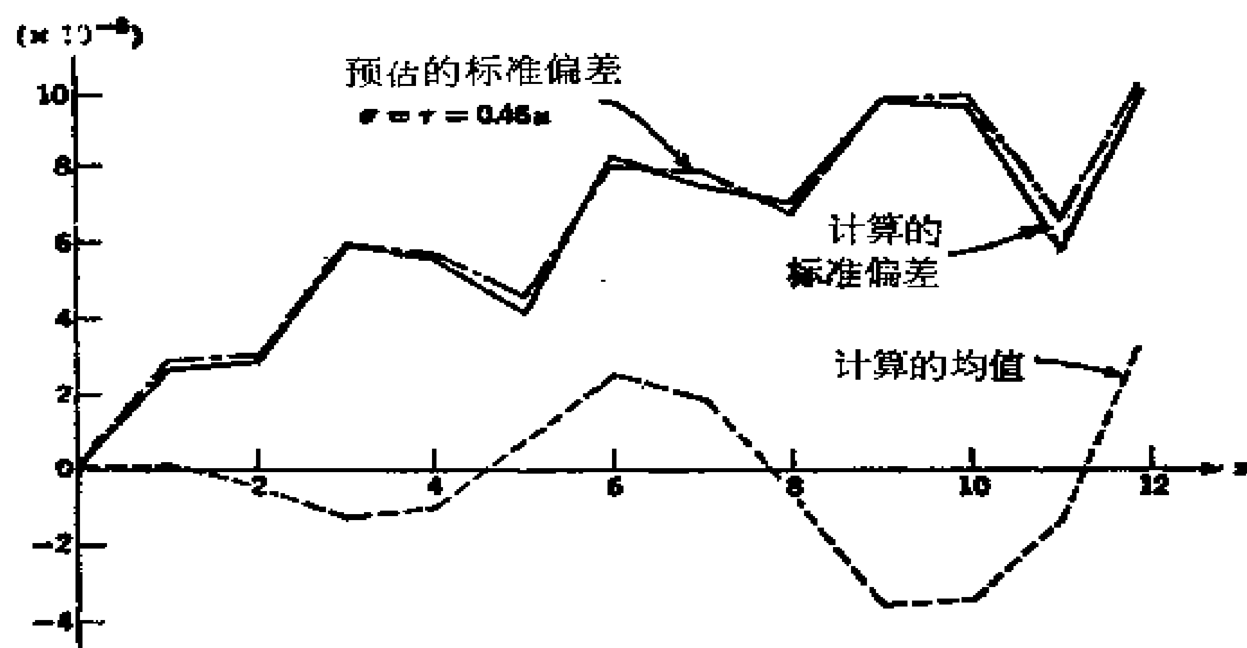


图 6.2　求和形式

是以双倍位精度进行计算获得的．舍入误差 $r_{n,q}=\tilde{y}_{n,q}-y_{n,q}$ 取样点为 $x_n=1,2,\cdots,12$．均值与标准偏差如图 6.1 和 6.2 所示．对于 $x_n=12$ 的值列于表 6.9 中．这些结果清楚地说明对于这个例子求和形式的优越性．

为了从理论上对这些结果进行分析，我们利用§3.4-8(iv)

表 6.9

	均 值	标准偏差
传统形式	-1×10^{-6}	8×10^{-6}
求和形式	4×10^{-8}	10×10^{-8}

中提出的理论. 为此假设局部舍入误差 ε_m 及 η_m 为对称分布，独立随机变量具有方差

$$\operatorname{var}(\varepsilon_m) = \frac{1}{3}\theta^2 u^2 y_m^2,$$
$$\operatorname{var}(\eta_m) = \frac{1}{3}\theta^2 u^2 F_m^2, \tag{6-168}$$

其中 θ 为一常数. 由于 IBM709 计算机固定地以底 $b=2$ 进行计算，从 (3-167) 我们得到 $\frac{1}{2}\leqslant\theta\leqslant 1$，而且还知道

$$u = 2^{-27}.$$

由于量 y_m 近似 $y(x_m)=\cos x_m$，对于传统的形式，定理 6.10 预估出

$$\operatorname{var}(r_n) = \frac{\theta^2 u^2}{3h^2}\{v(x_n)+O(h)\},$$

其中 $v(x)$ 为初值问题

$$v''' = -4v' + 2(\cos x)^2,\quad v(0)=v'(0)=v''(0)=0$$

的解. 积分便得

$$v(x) = \frac{4x - 2x\cos 2x - \sin 2x}{16}.$$

所得的值同时标在图 6.1 中. 常数 θ 是事后选择的，它为了使实验值与理论值尽可能地符合. 结果表明由 $\theta=0.79$ 所得的结果，符合得最好.

对于求和的格式，由于 F_m 近似 $y'(x_m)=-\sin x_m$，由定理 6.12 可期望

$$\mathrm{var}(r_n) = \frac{\theta^4 u^2}{3h^2}\{v(x_n) + t(x_n) + O(h)\},$$

其中 $v(x)$和 $t(x)$为初值问题

$$v''' = -4v' + 2(\sin x)^2,\ v(0) = v'(0) = v''(0) = 0,$$

$$t''' = -4t' + 2(\sin x)^2,\ t(0) = 0,\ t'(0) = 1,\ t''(0) = 0$$

的解. 积分得到

$$v(x) + t(x) = \frac{4x + 2x\cos x + \sin 2x}{8}.$$

所得的值标于图 6.2 中. 同样，为了符合得最好，还是选择 $\theta = 0.79$. θ 取这个值，事实上在所考虑的整个 x_n 的范围中理论值与实验值都符合得十分好.

6.5. 问题及附注

§6.1

1. 开始值. 从理论的观点来看，用 $x_0 = a$ 便可以开始计算. 在实际中，使用“虚”点 $x_n (n < 0)$ 往往是方便的. 例如假设 $y(x)$ 为满足 $y'(0) = 0$ 的方程 $y'' = f(x)y$ 的解，使用 Cowell 方法，其中 $f(x) = f(-x)$，那么由 $y_{-1} = y$ 以及从 (6-14) 当 $p = 1$ 时所满足的条件得到

$$y_1 = y_0 \frac{1 + \frac{5}{12} h^2 f_0}{1 - \frac{1}{12} h^2 f_1}.$$

证明对于这个 y_1 值，$y_1 - y(x_1) = O(h^6)$，当 $h \to 0$ 时.

2. 用步长 $h = 1$ 的 Cowell 方法 (6-14)，解初值问题

$$y'' + \frac{4x^2}{1 + x^2} y = 0,\ y(0) = 1,\ y'(0) = 0.$$

对于大的 x 值，$y(x) \sim A\cos(2x + \delta)$，其中 A 和 δ 为常数.

试确定它们.

3. Riccati 变换. 若 $y''=f(x)y$，且若函数 $u=y'/y$ 和 $v=y/y'$ 都是可微的，则它们分别满足一阶方程

$$u'=f(x)-u^2,\quad v'=1-f(x)v^2.$$

4. 构造一个显式，取 $p=6$，$k=4$ 的不稳定预估公式.

[提示：确定常数 α，β，γ 及 δ，使算子

$$y_{n+2}+y_{n-2}+\alpha(y_{n+1}+y_{n-1})+2\beta y_n$$
$$=h^2\{\gamma(f_{n+1}+f_{n-1})+2\delta f_n\}$$

为 6 阶. 有唯一的解为 $\alpha=16$，$2\beta=-14$，$\gamma=\frac{8}{3}$ 以及 $2\delta=\frac{44}{3}$]

5. 确定常数 α 和 β，使算子

$$y_{n+3}-y_{n+2}-y_{n+1}+y_n=h^2\{\alpha(f_{n+1}+f_{n+3})$$
$$+\beta(f_{n+1}+f_{n+2})\}$$

为 4 阶，并计算误差常数值.

6. (a) 证明与 Störmer 公式有关的算子当 $q=2$ 和 $q=3$ 时其阶数是一样的；

(b) 证明与多项式 (6-75) 有关的 6 阶算子当 $0<\varphi\leqslant\pi$ 时阶数不变[在 x_{n+2} 处展开].

7. 推广到高阶方程. 显然 §6.1 中的很多理论都可以推广到形如

$$y^{(q)}=f(x,y)$$

的微分方程，其中 q 为一个任意正整数. 然而定理 6.4 及 6.5 的证明实质上依赖于由

$$\left\{\frac{z}{\log\frac{1+z}{1-z}}\right\}^q=c_0^{(q)}+c_2^{(q)}z^2+c_4^{(q)}z^4+\cdots$$

确定的系数 $c_{2l}^{(q)}$，当 $q\leqslant 2$ 和 $l=1,2,\cdots$时，满足 $c_{2l}^q<0$.

证明当 $q>2.6$ 时 $c_4^{(q)}>0$，因此这些定理不能推广到 $q\geqslant 3$。

8.（继续）使用表达式

$$\left\{\frac{2x}{\log\frac{1+x}{1-x}}\right\}^3=1-x^2+\frac{1}{15}x^4+\frac{1}{945}x^6-\frac{16}{4725}x^8+\cdots$$

构造关于对微分方程

$$y'''=f(x,y)$$

积分的算子，它是稳定的（在适当的意义下）且在 $k=6$ 时 $p=10$. [与之有关的多项式 $r(x)$，除一个乘数因子外为唯一的：

$r(x)=x^3+\frac{5}{16}x^5$；见 Dahlquist [1959], p.30]

9. 确定与多项式

$$\rho(\zeta)=\zeta^6-2\zeta^3+1$$

有关的最佳算子.

§ 6.2

10. 确定与多项式

(a) $\rho(\zeta)=\zeta^4-2\zeta^2+1$,

(b) $\rho(\zeta)=\zeta^6-\zeta^4-\zeta^2+1$

有关的最佳方法的增长因子.

11. 计算与上题中的多项式相关联的多项式 $\rho_\nu(\zeta)$ 以及 Δ_ν 这个量.

12. 对在问题 10 中所考虑的多项式，确定满足引理 6.2 结论的常数 Γ 和 γ.

13. 讨论对初值问题

$$y''=\frac{8y^2}{1+2x},\ y(0)=1,\ y'(0)=-2$$

以一个具有误差常数 C 的 p 阶方法，所得数值解离散误差的渐近性态，[精确解：$y(x)=(1+2x)^{-1}$] 当

(a) $d=1$，而开始值为准确的，

(b) $d=2$，$\mu_3^2=\mu_4^2=-1$，而开始值由 p 阶的 Taylor 多项式得到.

14.* 对于特殊初值问题 $y''=A^2y$，$y(0)=1$，$y'(0)=A$ ($A=\text{const}$)，用形如 §5.3-1 中在某种意义上的直接计算法来证明定理 6.8 的结论.

15.* 对于初值问题 $y''=(p+1)(p+2)x^{-2}y$，$y(1)=1$，$y'(1)=p+2$，重复问题 14，其中 p 为方法的阶.

16.* 对写成下面形式：

$$\begin{aligned}\alpha_s y_{n+s}+\alpha_{s-1}y_{n+s-1}+\cdots+\alpha_{-s}y_{n-s}\\=h^2\{\beta_s f_{n+s}+\cdots+\beta_{-s}f_{n-s}\},\end{aligned}$$

其中 $\alpha_\mu=\alpha_{-\mu}$，$\beta_\mu=\beta_{-\mu}$，$\mu=0,1,\cdots,s$ 的对称差分方程 (6-23) 导出类似于第五章问题 32，33 及 34 的那些结果.

§6.3

17. 关于 $m_k(x)$的一个下界. 证明 $m_k(x)\geqslant|n_k(x)|$，其中 $n_k(x)$由

$$n_k''(x)=\mu_{2k}^2 g(x)n_k(x)+p(x),\quad n_k(a)=n_k'(a)=0$$

确定. $\left[\text{利用} \int_a^x p(t)|f(x,t)|dt\geqslant\left|\int_a^x p(t)f(x,t)dt\right|\right]$

18. 关于 $m_k(x)$的一个上界. 利用 Schwarz 不等式，证明

$$m_k^2(x)\leqslant v_k(x)\int_a^x p^2(t)q^{-1}(t)dt.$$

19. 若 ζ_μ 为一个稳定多项式 $\rho(\zeta)$ 的重根，证明

$$|\rho''(\zeta_\mu)|\leqslant|\alpha_k|\cdot 2^{k-1}.$$

$\left[\text{利用表达式 } \rho(\zeta)=\alpha_k(\zeta-\zeta_\mu)^2\prod_{\alpha\neq\mu}(\zeta-\zeta_\alpha)\right]$

20. 对于

$$y''=\frac{8y^2}{1+2x},\quad y(0)=1,\quad y'(0)=-2$$

的数值积分确定函数 $v_1(x)$。当 (a) $q=1$，(b) $q=y^2$ 时。

[答．令 $\zeta=1+2x$，便得

$$(a)\ v(x)=\frac{1}{136}\left\{-\frac{17}{3}\zeta^3+\zeta-\frac{\zeta^{1+\sqrt{17}}}{2-\sqrt{17}}-\frac{\zeta^{1-\sqrt{17}}}{2+\sqrt{17}}\right\},$$

$$(b)\ v(x)=\frac{1}{136}\left\{\frac{\zeta^{1+\sqrt{17}}}{\sqrt{17}}-\frac{\zeta^{1-\sqrt{17}}}{\sqrt{17}}-2\zeta\log\zeta\right\}]$$

21. 对方程

$$y''=\frac{8y}{(1+2x)^2},\ y(0)=1,\ y'(0)=-2$$

重复问题 20。

22. (a) 证明与定理 5.17 类似的如下结论：每一个稳定的、与一个相容算子有关的首项系数为 1 的多项式 $\rho(\zeta)$ 具有形式 $\zeta^m\Pi$，其中 Π 为割圆多项式的乘积，在乘积里 $\Phi_1(\zeta)$ 恰好出现二次而每个其它的因子至多二次．(b) 列举所有具有适合的相容性和 $m=0$，$k=6$ 的稳定的，首项系数为 1 的多项式。

§ 6.4

23. 利用 Cowell 方法的"求和"形式，重复问题 2。

24. 对问题 20 所讨论的情形，确定函数 $t(x)$。

25*. 如果用 $B_0, B_1, B_2, \cdots$ 表示 Bernoulli 数，如 Hildebrand [1956]，p. 150 中所定义，证明由 (6-145) 所定义的开始常数 H 满足

$$\alpha H=B_0y'(a)+\frac{B_1}{1!}y''(a)h+\cdots+\frac{B_{p-1}}{(p-1)!}y^{(p)}(a)h^{p-1}$$
$$+\left(\frac{B_p}{p!}+C\right)y^{(p+1)}(a)h^p+O(h^{p+1}),\qquad(6\text{-}169)$$

其中 p 及 C 分别表示方法的阶与误差常数。

26.* 令 $\gamma_\nu^*(\nu = 0,1,2,\cdots)$ 为由(5-20)所定义的常数，若 $p \geqslant k+1$，证明

$$\alpha H = y'(a) + h\{\gamma_1^* f_0 - \gamma_2^* \nabla f_1 + \cdots + (-1)^{k-1}\gamma_k^* \nabla^{k-1} f_{k-1}\} + O(h^{k+1}). \quad (6\text{-}170)$$

注

§6.1-1. Cowell 和 Commelin [1910] 在它们的 Halley 彗星轨道的计算中，曾使用 $q = 4$ 的 (6-10) 其系数等同于我们的 σ_m 和 σ_m^*，它由 Bennett, Milne 以及 Bateman [1956], p. 82—83 所给出．然而对应于我们的 σ_6 和 σ_5^* 的值是有错误的．其它的直接积分二阶方程的特殊方法为 Zadiraka [1951], Jacobsen [1952], Mikelazde [1953a], Urabe 和 Yanagihara [1954], Sconzo [1954], de Vogelaere [1955], Salzer[1957] 所给出．

§6.1-8. 定理 6.4 和 6.5 中所包含的结果来自 Dahlquist [1959]，其中有一些为 Conte [1958] 独立地获得．

§6.2-2. 关于特殊方法的先验界为 Muhin [1952a], Weissinger [1952], Matthieu [1953], Serrais [1956], Sheldon, Zondek 和 Friedman [1957], Uhlmann [1957a] 所阐述．一般线性多步方法是由 Dahlquist [1959] 论述的．

§6.2-3. 弱稳定的图形讨论为 Collatz [1960], p. 136 所给出．

§6.4. Cowell 方法的求和形式与一个积分方法是密切相关的，该方法天文学者至少从 1800 年就已经知道，并称它为“二次和方法”或是“$\sum^2$ 方法”（见 von Oppolzer [1880], Jackson[1924]）．该方法为 Herrick[1951] 公布而引起一般计算公众的注意．然而，显然没有认识到该方法在代数上是等价于我们这里的 Cowell 方法，它的主要的优点在于减少舍入误差并且求和的好处不只局限于 Cowell 方法．求和形式在

一个非天文学的问题中 Cowell 方法的优点曾为 Dettmar 和 Schlüter [1958] 所指出。 这一节的理论结果 Henrici [1960] 曾简短地发表过，而附加的数值试验在 Henrici [1961] 的工作中报导过。

第III部分 边值问题

到目前为止，所讨论的方法主要是为了解初值问题而设计的．在这些问题中，所有为确定解的必要的附加条件(除微分方程外)都是在同一个点上给出的．然而在许多实际问题中，(附加条件)却在几个不同的点上给出．例如，想要确定一个从地球表面的一个指定的点在给定的时间内飞行到另一个点的弹道导弹的轨道．这种问题就称为边值问题[1]．由于对单个的一阶微分方程通常仅需一个附加条件就足以确定解，所以真正的边值问题至少涉及二阶的微分方程（或至少含有二个方程的一阶方程组）．

由条件

$$y''=f(x,y),\ y(a)=A,\ y(b)=B$$

所表示的边值问题或许是最简单的，其中 $b>a$ 且 A 和 B 为已知常数．将求解边值问题化为求解一系列的初值问题，理论上总是可能的．令 $y(x,\alpha)$ 表示将上述问题的条件 $y(b)$ 换成 $y'(a)=\alpha$ 所得到的初值问题的解，其中 α 是一个参数．因此上述边值问题等价于对 α 求解方程(一般是非线性)

$$y(b,\alpha)=B.$$

这可以通过任一个标准方法例如试位法或 Newton 法来实现．后面的方法是适用的，这是由于函数 $\eta(x)=y_\alpha(x,\alpha)$ 为初值问题（见 §7.1-1）$\eta''=f_y(x,y(x,\alpha))\ \eta,\ \eta(a)=0,\ \eta'(a)=1$ 的解．

1) 对于附加条件精确地在两个不同点上提出的，通常称为两点边值问题．

每求函数 $y(x,\alpha)$ 的值一次[如果使用 Newton 法则需求值 $\eta(x,\alpha)$] 要求解一个初值问题. 不过上述的"打靶"方法[在苏联也称为"drive-through"方法]也许还是提供了一个可行的方法. 在其它的情形下, 特别当微分方程组复杂的或当初值问题预示着数学的不稳定性时, 此时用其它的更为直接的方法处理已知问题可能更好一些. 在第七章中, 我们讨论一种在实际中常用的方法, 这种方法在于用一组隐式差分方程来代替给定的问题.

即使对于上面所考虑的简单的边值问题都可能发生有无穷多个解的情形——正如在问题 $y''+\pi^2y=0$, $y(0)=0$, $y(1)=0$ 中, 对任意的 C, $y(x)=C\sin\pi x$ 都是一个解——或者又如在问题 $y''+\pi^2y=0$, $y(0)=0$, $y(1)=1$ 中其解就不存在.

这些简单的例子指出, 边值问题的数学理论要比初值问题的理论复杂得多. 因此, 必须预料到这类问题在数值处理的理论上也有同样的问题. 为了能体现数学的连贯性同时篇幅又不太冗长, 在第七章中我们将把注意力局限于一类略为特殊的、但仍为非线性的问题. 这类问题所展示的大多数性质, 它们对更复杂的问题也是典型的. 此外, 这里将讨论的大多数方法都能适用于更一般的情形. 关于边值问题数值解法的更广泛的讨论, 读者可以参看 Collatz[1949], Collatz[1960], Fox [1957].

第七章　一类二阶非线性边值问题的直接方法

一个边值问题被称为M类的，如果它具有形式

$$y'' = f(x,y),\ y(a) = A,\ y(b) = B, \qquad (7-1)$$

其中$-\infty < a < b < \infty$，A和B为任意的常数，而且函数$f(x,y)$除了满足存在定理1.1中的条件外，再加上$f_y(x,y)$为连续且满足

$$f_y(x,y) \geqslant 0,\ a \leqslant x \leqslant b,\ -\infty < y < \infty \qquad (7-2)$$

的条件．本章的理论研究仅涉及到M类问题．某些推广将在本章末的问题中被涉及到．

7.1．求解的方法

7.1-1．*唯一解的存在性*．将问题集中到M类的主要目的是为了下面的定理：

定理7.1．　M类的边值问题有唯一的解．

证．令$y(x,\alpha)$表示初值问题$y'' = f(x,y)$，$y(a) = A$，$y'(a) = \alpha$的解．利用微分方程的一个标准定理[1]，$y(x,\alpha)$为x和α的一个连续函数，而且$y_\alpha(x,\alpha)$也存在，它对于$x \in [a,b]$及所有的α是连续的．为了证明关于α的方程

$$y(b,\alpha) = B$$

确实有一个解，我们将证明

1) 例如，参看 Coddington and Levinson [1955] 第一章，定理7.5．

$$y_\alpha(b,\alpha) \geqslant b - a. \tag{7-3}$$

由对所有 α 值都有定义的单调函数的导数是大于零的，假设取每一个值只能有一次，便得到所要求的结论.

为了证明 (7-3)，我们将恒等式 $y''(x,\alpha) = f(x, y(x,\alpha))$ 对 α 微分. 记 $\eta(x) = y_\alpha(x,\alpha)$，便得到

$$\eta''(x) = f_y(x, y(x,\alpha))\eta(x), \quad a \leqslant x \leqslant b. \tag{7-4}$$

从 $y(x,\alpha)$ 的定义，我们有

$$\eta(a) = 0, \quad \eta'(a) = 1. \tag{7-5}$$

我们将证明当 $a \leqslant x \leqslant b$ 时 $\eta(x) \geqslant x - a$. 假设对于某个 $\zeta \in [a,b]$，$\eta(\zeta) < \zeta - a$. 当 $x - a$ 为小的正值时，由于 $\eta(x) > 0$，不失一般性，我们可以假设

$$\eta(x) > 0, \quad 当\ a < x \leqslant \zeta. \tag{7-6}$$

利用中值定理，$\eta(\zeta) = (\zeta - a)\eta'(\zeta_1)$，对于某个 $\zeta_1 \in (a,\zeta)$. 由此得 $\eta'(\zeta_1) < 1$. 将中值定理应用于函数 $\eta'(x)$，对某个 $\zeta_2 \in (a,\zeta_1)$，我们得到 $\eta'(\zeta_1) - \eta'(a) = (\zeta_1 - a)\eta''(\zeta_2)$. 由于 $\eta'(a) = 1$，于是我们有 $\eta''(\zeta_2) < 0$. 这就与微分方程 (7-4) 相矛盾，因为 $f_y \geqslant 0$ 和 (7-6). 由此便得 $\eta(x) \geqslant x - a$. 所要求的关系式 (7-3) 即为 $x = b$ 的特殊情形.

7.1-2. *建立一个有限差分格式*. 对于 M 类的边值问题的直接数值解，我们引进点 $x_n = a + nh (n = 0, 1, \cdots, N)$，其中 $h = (b - a)N^{-1}$，而 N 为一个适当的整数. 于是就设计了一个差分格式来确定 y_n，期望它在点 x_n 上近似精确解 $y(x_n)$. 得到这样一个差分格式的自然途径，是要求 y_n 在每一个内网格点 x_n 上满足一个类似于 (6-23) 的差分方程

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n - h^2\{\beta_k f_{n+k} + \cdots + \beta_0 f_n\} = 0. \tag{7-7}$$

按照这样的方法，再选取这些系数，使得相关的差分算子 (6-24) 对 $y'' = f(x, y)$ 的解来说为小量. 这样不再需要假设

$\alpha_k \neq 0$，因为由此产生的关于 y_n 的方程组在任何情形下都是隐式的，因而(7-7)不再需要对 y_{n+k} 来说是可解的．然而，有另外的困难，正如在任何代数问题中那样，我们为了确定未知量需要有与未知量个数一样多的方程．由于 y_0 和 y_N 是由边界条件所确定的，在我们这种情形，未知量为 $y_1, \cdots, y_{N-1}$．如果差分表达式的步数 >2，就要引进新的未知量，例如 y_{-1} 或 y_{N+1}，而对它们并没有立出方程．

这个困难通过在边界点附近处将差分方程适当地加以修改便可克服；如果 $k=2$ 这个最小的可能值，全然不会产生这种情形．由 §6.1 推得，如果在(7-7)中 $k=2$，又如果有关的差分算子有正的阶 p，那么差分方程必与形如[1]

$$-y_{n-1} + 2y_n - y_{n+1} + h^2\{\beta_0 f_{n-1} + \beta_1 f_n + \beta_2 f_{n+1}\} = 0 \tag{7-8}$$

的方程成比例，其中 $\beta_0 + \beta_1 + \beta_2 = 1$．关于边值问题，最常用的差分方程为

$$-y_{n-1} + 2y_n - y_{n+1} + h^2 f_n = 0 \ (p = 2) \tag{7-9}$$

以及

$$-y_{n-1} + 2y_n - y_{n+1} + \frac{1}{12}h^2(f_{n-1} + 10f_n + f_{n+1}) = 0 \quad (p = 4). \tag{7-10}$$

看来对于大多数实际问题，建立在(7-8)上的差分格式是适合的．因此，在本章中我们将只涉及形如(7-8)的差分方程．关于更复杂的差分格式的构造，读者可参看 Fox [1957]．

以下两小节中的注释并非以后理论研究的课题．

第二类和第三类边界条件．附加在(7-1)中的边界条件，称为第一类边界条件，在实际中也出现形如

$$\alpha y(a) + \beta y'(a) = A, \quad \gamma y(b) + \delta y'(b) = B \tag{7-11}$$

的条件，其中 $\alpha, \beta, \gamma, \delta$ 为常数，$\beta^2 + \delta^2 > 0$．条件(7-11)

1) 为了以后方便，选取在这个表达式中的记号和下标．

被称为第二类或第三类的边值条件是按照 $\alpha^2+\gamma^2=0$ 或 >0 而定的。上述格式容易推广到这些情形。在(7-8)中令 $n=0$，并用(7-11)的第一个方程可消去 y_{-1}，便得到关于 y_0 的一个附加方程，其中 $y(a)=y_0$ 及

$$y'(a)=(y_1-y_{-1})(2h)^{-1}.$$

为了得到 y_N 的方程，可以采用类似的方法．关于第三类情形的某些问题解的存在性，可参看问题 8.

变步长．在许多问题中，特别是解 $y(x)$ 必须预想到在区间 $[a,b]$ 的某些部份变化快而在另一些部分变化慢，这就没有理由在整个区间使用同一个步长 h．看来一个较为自然的方法是将区间 $[a,b]$ 分解成一系列的子区间，而且在每个子区间中，按照预料到的 $y(x)$ 的性态，而以不同的 h 求解．上述的差分格式是容易适合于这种情形的，如果对任意二个相邻区间来说，在一个区间中所用步长为另一个区间中所用的整数倍．对于二个步长的比为 2:1 的情形，如图 7.1 所示．

7.1-3. *差分格式的解；线性情形*．与第三章的情形一样，下面的讨论可用矩阵和向量记号加以简化．引入向量[1]

$$\mathbf{Y}=\begin{pmatrix} y_1\\ y_2\\ \vdots\\ y_{N-1}\end{pmatrix},\quad \mathbf{f}(\mathbf{Y})=\begin{pmatrix} f(x_1,y_1)\\ f(x_2,y_2)\\ \vdots\\ f(x_{N-1},y_{N-1})\end{pmatrix},$$

$$\mathbf{a}=\begin{pmatrix} A-\beta_0h^2f(x_0,A)\\ 0\\ \vdots\\ 0\\ B-\beta_2h^2f(x_N,B)\end{pmatrix}$$

1) 必须指出，向量 $\mathbf{Y}$ 的元素为近似于在不同点上同一个函数的那些值，而在第三章中 $\mathbf{Y}$ 的元素为近似在同一点上的不同函数那些值．

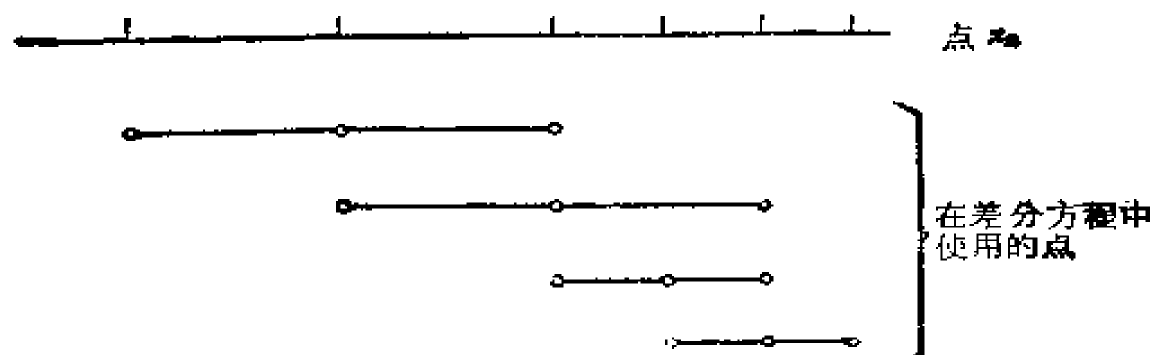


图 7.1 变步长

以及矩阵

$$\mathbf{J}=\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix},$$

$$\mathbf{B}=\begin{pmatrix} \beta_1 & \beta_2 & & & \\ \beta_0 & \beta_1 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_0 & \beta_1 & \beta_2 \\ & & & \beta_0 & \beta_1 \end{pmatrix}$$

(在这些矩阵中所有的不在主对角线上或次对角线上的元素均为零). 从要求 (7-7) 对 $n=1, 2, \cdots N-1$ 成立所给出的方程组,可以写成紧凑的格式

$$\mathbf{JY}+h^2\mathbf{Bf}(\mathbf{Y})=\mathbf{a}. \tag{7-12}$$

在一般的情形下,求解这个方程组的方法将在 § 7.1-4 中讨论. 在这一节中, 我们的注意力集中在给定的微分方程为

线性的情形，即 $f(x,y)$ 具有形式

$$f(x,y)=g(x)y+k(x),$$

其中 $g(x)$ 及 $k(x)$ 为给定的函数。定义对角阵

$$\mathbf{G}=\begin{pmatrix} g(x_1) & & & \\ & g(x_2) & & \\ & & \ddots & \\ & & & g(x_{N-1}) \end{pmatrix}$$

和向量

$$\mathbf{k}=\begin{pmatrix} k(x_1) \\ k(x_2) \\ \vdots \\ k(x_{N-1}) \end{pmatrix},$$

我们可以写成

$$\mathbf{f}(\mathbf{y})=\mathbf{G}\mathbf{y}+\mathbf{k}.$$

那么方程组 (7-12) 化成线性方程组

$$\mathbf{A}\mathbf{y}=\mathbf{b}, \tag{7-13}$$

其中

$$\mathbf{A}=\mathbf{J}+h^2\mathbf{B}\mathbf{G},\ \mathbf{b}=\mathbf{a}-h^2\mathbf{B}\mathbf{k}. \tag{7-14}$$

虽然线性方程组 (7-13) 的阶可以很大(事实上，当 $h\to 0$ 时阶趋于无穷)，而它的数值解却相对地容易求出，这是由于矩阵 $\mathbf{A}$ 包含有矩阵 $\mathbf{J}$ 和 $\mathbf{B}$，而它们却具有仅在主对角线及其相邻的次对角线上有非零元素的性质。这种类型的矩阵称为三对角的。 事实上，如果 $\mathbf{A}=(a_{nm})$，那么由 (7-14)，我们有

$$a_{n,n}=2+h^2\beta_1 g_n,\quad a_{n,n-1}=-1+h^2\beta_0 g_{n-1},$$

$$a_{n,n+1}=-1+h^2\beta_2 g_{n+1},\quad n=1,2,\cdots,N-1$$

且当 $|n-m|>1$ 时 $a_{n,m}=0$。

含有非奇异的三对角矩阵的线性方程组采用 Gauss 算法最容易求解，它可以方便地概括如下。假如我们已经成功地确定了二个具有特殊形式的非奇异矩阵 $\mathbf{L}=(l_{mn})$ 和 $\mathbf{U}=(u_{mn})$：

$$\mathbf{L}=\begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & 0 & \\ & l_{32} & 1 & & \\ & & \ddots & \ddots & \\ 0 & & & l_{N-1,N-2} & 1 \end{pmatrix}$$

$(l_{mm}=1; l_{mn}=0$ 当 $n>m$ 或是 $n<m-1$ 时)

$$\mathbf{U}=\begin{pmatrix} u_{11} & u_{12} & & & \\ & u_{22} & u_{23} & & \\ & & \ddots & \ddots & \\ & & & \ddots & u_{N-2,N-1} \\ & & & & u_{N-1,N-1} \end{pmatrix}$$

$(u_{mn}=0$ 当 $n<m$ 或 $n>m+1)$

具有性质

$$\mathbf{LU}=\mathbf{A}. \tag{7-15}$$

为了求解 (7-13)，我们首先确定向量 $\mathbf{z}$，使得

$$\mathbf{Lz}=\mathbf{b}, \tag{7-16}$$

然后确定 $\mathbf{y}$ 使

$$\mathbf{Uy}=\mathbf{z}. \tag{7-17}$$

由于 $\mathbf{y}=\mathbf{U}^{-1}\mathbf{z}=\mathbf{U}^{-1}\mathbf{L}^{-1}\mathbf{b}=\mathbf{A}^{-1}\mathbf{b}$，如此确定的向量 $\mathbf{y}$ 满足 (7-13)。

为了实现这个方案，我们指出 (7-15) 等价于关系式

$$u_{11}=a_{11}, \tag{7-18}$$

$$l_{n,n-1}u_{n-1,n-1}=a_{n,n-1}, n=2,3,\cdots,N-1, \tag{7-19a}$$

$$l_{n,n-1}u_{n-1,n}+u_{n,n}=a_{n,n},\tag{7-19b}$$

$$u_{n,n+1}=a_{n,n+1},\quad n=1,2,\cdots,N-2.\tag{7-19c}$$

关系式(7-19c)立即给出 $u_{n,n+1}$；为了递推地求得 $l_{n,n-1}$ 及 $u_{n,n}$，可以重新将关系式(7-19a)和(7-19b)排列如下：

$$u_{11}=a_{11},\tag{7-20}$$

$$l_{n,n-1}=\frac{a_{n,n-1}}{u_{n-1,n-1}},\quad n=2,\cdots,N-1,\tag{7-21a}$$

$$u_{n,n}=a_{n,n}-l_{n,n-1}a_{n-1,n}.\tag{7-21b}$$

当 $u_{n-1,n-1}=0$ 时算法(7-21)失效. 如果有这种情形，那么，用 $\det\mathbf{A}$ 表示 $\mathbf{A}$ 矩阵的行列式，我们有

$$\det\mathbf{A}=\det\mathbf{U}\det\mathbf{L}=u_{11}u_{22}\cdots u_{N-1,N-1}=0,$$

于是 $\mathbf{A}$ 为奇异矩阵.

向量 $\mathbf{z}$ 与 $\mathbf{L}$ 可由关系式

$$z_1=b_1$$

$$l_{n,n-1}z_{n-1}+z_n=b_n,\quad n=2,\cdots,N-1$$

同时确定，它可以递推地改写成形式

$$z_1=b_1,\tag{7-22a}$$

$$z_n=b_n-l_{n,n-1}z_{n-1},\quad n=2,\cdots,N-1.\tag{7-22b}$$

如果将(7-17)书写成分量形式，我们得到

$$u_{N-1,N-1}y_{N-1}=z_{N-1},$$

$$u_{n,n}y_n+u_{n,n+1}y_{n+1}=z_n,\quad n=1,2,\cdots,N-2.$$

为了从最后一个分量开始求得 $\mathbf{y}$ 的分量，这些关系式可改写成如下的形式，于是便得到

$$y_{N-1}=\frac{z_{N-1}}{u_{N-1,N-1}},\tag{7-23a}$$

$$y_n=\frac{z_n-u_{n,n+1}y_{n+1}}{u_{n,n}},\ n=N-2,\cdots,1.\tag{7-23b}$$

完整的过程可以概括如下：

(I) 由 $n=1$ 开始，从以下关系式计算 u_{nn} 和 z_n:

$$\boxed{\begin{aligned} & u_{11}=a_{11},\quad z_1=b_1, \\ & l_{n,n-1}=\frac{a_{n,n-1}}{u_{n-1,n-1}}, \\ & u_{n,n}=a_{n,n}-l_{n,n-1}a_{n-1,n} \qquad n=2,3,\cdots,N-1. \\ & z_n=b_n-l_{n,n-1}z_{n-1}, \end{aligned}}$$

量 $l_{n,n-1}$ 无需保存，除去同样的方程组对不同的非齐次项重复求解而外。 量 $a_{n,n}$ 和 b_n 在计算出 $u_{n,n}$ 和 z_n 后便无需再用。

(II) 由 $n=N-1$ 开始，从以下关系式计算 y_n:

$$\boxed{\begin{aligned} & y_{N-1}=\frac{z_{N-1}}{u_{N-1,N-1}}, \\ & y_n=\frac{z_n-a_{n,n+1}y_{n+1}}{u_{n,n}},\ n=N-2,\cdots,1. \end{aligned}}$$

整个过程大约需要 $3N$ 次加法，$3N$ 次乘法和 $2N$ 次除法。 这与没有零元素的 N 阶矩阵的方程组求解仅乘法就必须完成 $1/3N^3$ 次相比，就十分有利了。

数值例子。 求解由边值问题 $y''=0$, $y(0)=0$, $y(1)=1$ 所产生的三对角线方程组，采用算子 (7-9), $h=0.2$。 表 7.1 中的箭头说明计算辅助量的顺序。

用 Gauss 算法求解简单边值问题

n	$a_{n,n-1}$	a_{nn}	$a_{n,n+1}$	b_n	$l_{n,n-1}$	u_{nn}	z_n	y_n
1		2	-1	0		2	0	$\frac{1}{5}$
2	-1	2	-1	0	$-\frac{1}{2}$	$\frac{3}{2}$	0	$\frac{2}{5}$
3	-1	2	-1	0	$-\frac{2}{3}$	$\frac{4}{3}$	0	$\frac{3}{5}$
4	-1	2		1	$-\frac{3}{4}$	$\frac{5}{4}$	1	$\frac{4}{5}$

上述算法仅为寻找 (7-13) 的解 $\mathbf{y}$ 服务的，假设这个解

是存在的话． M 类问题解的存在性（或是相当于 $\mathbf{A}$ 为非奇异)在 §7.2-1 中证明．

7.1-4. *差分格式的解；非线性情形．* 如果函数 $f(x,y)$ 对 y 是非线性的，便不能希望用代数方法求解方程组 (7-12)，必须依靠某些迭代方法．为此，我们所介绍的方法是一种推广的解超越方程组的 Newton-Raphson 方法．

在单个方程的情形下，Newton-Raphson 方法在于对已知方程 $f(x)=0$ 的线性化，通过将 $f(x)-f(x^{(0)})$ 换成在 $x^{(0)}$ 处 $f(x)$ 的微分．而 $x^{(0)}$ 被认为是接近真解的，然后解线性化的方程 $f(x^{(0)})+f'(x^{(0)})\Delta x=0$． 于是 $x^{(1)}=x^{(0)}+\Delta x$ 值被取作一个较好的近似值，如有必要可延续这一过程．完全类似地，如果认为 $\mathbf{y}^{(0)}$ 是一个接近方程

$$\mathbf{J}\mathbf{y}+h^2\mathbf{B}\mathbf{f}(\mathbf{y})-\mathbf{a}=0 \qquad (7\text{-}24)$$

真解的向量，于是剩余向量

$$\mathbf{r}(\mathbf{y}^{(0)})=\mathbf{J}\mathbf{y}^{(0)}+h^2\mathbf{B}\mathbf{f}(\mathbf{y}^{(0)})-\mathbf{a} \qquad (7\text{-}25)$$

为小量．我们将函数 $\mathbf{r}(\mathbf{y})$ 的增量换成在 $\mathbf{y}=\mathbf{y}^{(0)}$ 处它们的微分，并从所得的线性方程组中解出向量 $\mathbf{Y}$ 的增量，这个量我们记为 $\Delta\mathbf{y}$． 由于表达式 $\mathbf{J}\mathbf{y}$ 对 $\mathbf{y}$ 来说已是线性的，向量 $\mathbf{f}(\mathbf{y})$ 在 $\mathbf{y}=\mathbf{y}_0$ 处的微分为 $\mathbf{F}(\mathbf{y}^{(0)})\Delta\mathbf{y}$，其中 $\mathbf{F}(\mathbf{y})$ 表示对角矩阵：

$$\mathbf{F}(\mathbf{y})=\begin{pmatrix} f_y(x_1,y_1) & & & 0 \\ & f_y(x_2,y_2) & & \\ & & \ddots & \\ 0 & & & f_y(x_{N-1},y_{N-1}) \end{pmatrix}. \qquad (7\text{-}26)$$

因此线性化的方程组 (7-24) 可写成

$$\mathbf{r}(\mathbf{y}^{(0)})+(\mathbf{J}+h^2\mathbf{B}\mathbf{F}(\mathbf{y}^{(0)}))\Delta\mathbf{y}=\mathbf{0}, \qquad (7\text{-}27)$$

而它的解由

$$\Delta\mathbf{y}=\Delta\mathbf{y}^{(0)}=-\mathbf{A}(\mathbf{y}^{(0)})^{-1}\mathbf{r}(\mathbf{y}^{(0)})$$

给出，假如矩阵

$$\mathbf{A}(\mathbf{y}) = \mathbf{J} + h^2\mathbf{BF}(\mathbf{y}) \tag{7-28}$$

的逆矩阵在 $\mathbf{y} = \mathbf{y}^{(0)}$ 处存在的话。如果一切进行正常，向量 $\mathbf{y}^{(1)} = \mathbf{y}^{(0)} + \Delta\mathbf{y}^{(0)}$ 将是精确解的一个较好的近似，剩余向量 $\mathbf{r}(\mathbf{y}^{(1)})$ 将会更小，用 $\mathbf{y}^{(1)}$ 取代 $\mathbf{y}^{(0)}$ 的位置并重复这个过程直到收敛为止.

在 §7.2-4 中曾证明关于M类问题，对于充分小的 h 值，方程组 (7-24) 有唯一的解 $\mathbf{y}$，而且 Newton 方法产生的向量序列 $\mathbf{y}^{(n)} n = 1, 2, \cdots$ 可迅速地收敛于 $\mathbf{y}$，假如初始近似值 $\mathbf{y}^{(0)}$ 不太差的话. 在这一节中，我们主要考虑计算方法问题. 从这方面来说，应该指出解 (7-27) 是无需计算 $\mathbf{A}(\mathbf{y}^{(0)})$ 的逆矩阵，唯一需要解的是关于 $\Delta\mathbf{y}$ 的分量的线性方程组

$$\mathbf{A}(\mathbf{y}^{(0)})\Delta\mathbf{y} = -\mathbf{r}(\mathbf{y}^{(0)}).$$

利用矩阵 $\mathbf{A}(\mathbf{y}^{(0)})$ 同样为三对角的这一事实，便可大大简化求解过程. 事实上，如果 $\mathbf{A}(\mathbf{y}^{(0)}) = (a_{mn})$，我们有

$$a_{n,n-1} = -1 + h^2\beta_0 f_y(x_{n-1}, y_{n-1}^{(0)}),\quad n = 2, \cdots, N-1,$$

$$a_{n,n} = 2 + h^2\beta_1 f_y(x_n, y_n^{(0)}),\quad n = 1, \cdots, N-1,$$

$$a_{n,n+1} = -1 + h^2\beta_2 f_y(x_{n+1}, y_{n+1}^{(0)}),\quad n = 1, \cdots, N-2.$$

而其它的所有元素均为零. 因而在 §7.1-3 中描述的方法便可直接采用. Newton 方法在每进行一步时，除了与解线性方程组有关的工作外，只需外加对剩余向量 $\mathbf{r}(\mathbf{y}^{(0)})$ 和对偏导数 $f_y(x_n, y_n)(n = 1, 2, \cdots, N-1)$ 进行计算.

数值例子. 取 $h = 0.1$ 以差分算子 (7-9) 解由非线性边值问题

$$y'' = -2 + \sinh y,\quad y(0) = 0,\quad y(1) = 0 \tag{7-29}$$

所产生的差分方程. 正如所说，该问题包含了 9 个未知量 $y_1, y_2, \cdots, y_9$. 然而利用精确解 $y(x)$ 满足 $y(x) = y(1-x)$ 的事实，可使计算工作量几乎减少一半（它可由定理 6.1 得到

(7-29) 解的唯一性很快地得出]. 因此,我们可以假设

$$y_n = y_{10-n},$$

因而只需计算值 $y_1, \cdots, y_5$. 利用 $y_6 = y_4$, 由 (7-9) 产生的差分方程为

$$\begin{aligned}
2y_1 - y_2 + 0.01(-2 + \sinh y_1) &= 0, \\
-y_1 + 2y_2 - y_3 + 0.01(-2 + \sinh y_2) &= 0, \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
-y_3 + 2y_4 - y_5 + 0.01(-2 + \sinh y_4) &= 0, \\
-2y_4 + 2y_5 + 0.01(-2 + \sinh y_5) &= 0.
\end{aligned}$$

我们取函数 $y(x) = x(1-x)$ 的值作为一次近似，该函数是 (7-29) 略去非线性项 $\sinh y$ 得到的解，所产生的剩余向量为

$$\mathbf{r}(\mathbf{y}^{(0)}) = (0.00090122, 0.00160684, 0.00211547, 0.00242311, -0.00252612)$$

以及矩阵

$$\mathbf{A}(\mathbf{y}^{(0)}) = \begin{pmatrix} 2.01004053 & -1 & & & \\ -1 & 2.01012827 & -1 & & \\ & -1 & 2.01022131 & -1 & \\ & & -1 & 2.01078939 & \\ & & & -1 & 2.01031413 \end{pmatrix}$$

(未标出的元素为零). 求得校正量 $\Delta\mathbf{y}^{(0)}$ 为

$$\Delta\mathbf{y}^{(0)} = (-0.0075325, -0.0142394, -0.0194837, -0.0228116, -0.0239511).$$

新的近似值 $\mathbf{y}^{(1)} = \mathbf{y}^{(0)} + \Delta\mathbf{y}^{(0)}$ 产生的剩余向量为

$$\mathbf{r}(\mathbf{y}^{(1)}) = (0.0000000, 0.0000002, 0.0000003, 0.0000006, 0.0000008).$$

进一步应用此过程产生向量

$$\mathbf{y}^{(2)} = (0.0824662, 0.1457580, 0.190525, 0.2171837, 0.2260438),$$

对它来说按机器保留数字 $\mathbf{r}(\mathbf{y}^{(2)}) = 0$。

象现在面临的这种问题，其矩阵 $\mathbf{A}(\mathbf{y})$ 的元素并不很敏感地依赖于 $\mathbf{y}$，通常允许仅在 $\mathbf{y} = \mathbf{y}^{(0)}$ 处计算 $\mathbf{A}(\mathbf{y})$，而从关系式

$$\mathbf{A}(\mathbf{y}^{(0)})\Delta\mathbf{y}^{(\nu)} = -\mathbf{r}(\mathbf{y}^{(\nu)}), \quad \nu = 1, 2, \cdots$$

确定更好的近似值.

7.2. 差分方程解的存在性

在 7.2-1 及 7.2-2 中，我们将讨论与边值问题的数值解有关的矩阵的某些概念. 作为一个推论，我们将得到线性方程组 (7-12) 和 (7-27) 恒有唯一解的结果. 在 7.2-3 和 7.2-4 中，我们将对含有 n 个未知量的 n 个非线性方程的一般方程组的 Newton 方法进行讨论. 作为这个讨论的一个应用，我们将对 M 类边值问题，建立一个定理来保证形如在 7.1-4 中所讨论的 Newton 方法的收敛性.

7.2-1. 不可约矩阵。 令 W 为前 n 个整数所组成的集合，$W = \{1, 2, \cdots, n\}$。一个矩阵 $\mathbf{A} = (a_{ij})$ 称为可约的，如果 W 能分解成二个非空且不交的子集 S 和 T，使得当 $i \in S$ 和 $j \in T$ 时 $a_{ij} = 0$。象通常在集合论中那样来表示，用 $\cup$ 表示二个集合的和集，以 $\cap$ 表示二个集合的交集，并以 $\varnothing$ 表示空集. 对 S 和 T 提出的这个条件，可以形式化地记作 $S \cup T = W$，$S \cap T = \varnothing$，$S \neq \varnothing$，$T \neq \varnothing$.

通过将序列 $\{1, 2, 3, \cdots, n\}$ 进行以下方式的置换，使得前 $r(1 \leqslant r < n)$ 个元素属于 S 而后 $n - r$ 个元素属于 T.

可以看出，一个可约矩阵与一个形如

$$\begin{pmatrix} \mathbf{A}_{11} & 0 \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

的矩阵相似，其中 $\mathbf{A}_{11}$ 及 $\mathbf{A}_{22}$ 分别为 r 和 $n-r$ 阶的方阵，$\mathbf{A}_{21}$ 为一个 $(n-r)\times r$ 的长方阵，而 $\mathbf{0}$ 为 $r\times(n-r)$ 零矩阵。因此，一个具有可约性非奇异矩阵的线性方程组具有这样的性质，即方程组中的某些方程可以在不需考虑其它方程的情形下解出。

如果一个矩阵不是可约的，就称它为不可约，对我们来说，建立矩阵 (7-14) 的不可约性是重要的。就这方面而论，下面的定理是有用的。

定理 7.2. 一个 $n\geqslant 2$ 阶的矩阵 $\mathbf{A}=(a_{ij})$ 为不可约的充要条件是对于任意二个整数 i 和 j，$i\in W$，$j\in W$，存在着 $\mathbf{A}$ 的一串非零元素，其形如

$$\{a_{i,i_1}, a_{i_1,i_2}, a_{i_2,i_3}, \cdots, a_{i_{m-1},j}\}. \tag{7-30}$$

$\mathbf{A}$ 的形如 (7-30) 的非零元素串被称为一个链。如果

$$a_{ij}\neq 0,$$

链 (7-30) 可假设为由一个元素 a_{ij} 组成。

证. (i) 假设 $\mathbf{A}$ 为可约的，且令 S 和 T 是上面所介绍的集合。由于在链 $\{a_{i_{\nu-1},i_\nu}\}$ 中的所有元素均异于零，这只有三种可能性：

(a) $i_{\nu-1}\in T$，$i_\nu\in T$；

(b) $i_{\nu-1}\in T$，$i_\nu\in S$；

(c) $i_{\nu-1}\in S$，$i_\nu\in S$，$\nu=1,\cdots,m$。

换言之，定义一个链的下标 $(i_0,i_1,\cdots,i_m)$ 只能一次改变它们的所属关系到集合 S 和 T 中的一个，如果它们确实有变化的话，只能从 T 变到 S 中。由此得出不可能存在一个 $i=i_0\in S$，$j=i_n\in T$ 的链。

(ii) 现假设不存在 $i=i^*$, $j=j^*$ 的链. 令 S 为所有使得对 (i^*,j) 存在着一个链的下标 j 的集合. 如果 $S=\varnothing$,这意味着 $a_{i^*,j}=0$ 对 $j=1,\cdots,n$, 于是 $\mathbf{A}$ 为可约的. 如果 $S\neq\varnothing$, 令 $T=W-S$. 由于 $j^*\in T, T\neq\varnothing$, 显然 $T\cup S=W$ 且 $T\cap S=\varnothing$. 我们断言, 当 $i\in S$, $j\in T$ 时 $a_{ij}=0$ 蕴含着 $\mathbf{A}$ 为可约的. 如果对于 $i\in S$, $j\in T$, $a_{ij}\neq 0$,那么由于对(i^*,i)存在着一个链, 这个链便可以通过在它的右端加上一个链 (i,j) 而加以扩充. 这将意味着 j 在 S 中而不在 T 中,它便与 S 的定义相矛盾. 这就完成了定理 7.2 的证明.

定理 7.2 的推论. 一个三对角矩阵 $\mathbf{A}=(a_{ij})$ 为不可约,当且仅当

$$\begin{aligned} &a_{i,i-1}\neq 0,\ i=2,3,\cdots,n \\ \text{而}\quad &a_{i,i+1}\neq 0,\ i=1,2,\cdots,n-1. \end{aligned} \tag{7-31}$$

证. (i) 假设 (7-31) 成立,并令 (i,j) 为满足 $i\leqslant j-2$ 的一对下标,于是由序列 $\{a_{i,i+1}, a_{i+1,i+2},\cdots,a_{j-1,j}\}$ 给出了一个链. 对 $i\geqslant j+2$ 有类似的论断成立. 如 $a_{i,i}=0$, $\{a_{i,i+1}, a_{i+1,i}\}$ 为一个适合的链,由此得出 $\mathbf{A}$ 为不可约的.

(ii) 如果对某些 $m\in W$, $m>1$, $a_{m,m-1}=0$, 令

$$S=\{m,m+1,\cdots,n\},\ T=\{1,2,\cdots,m-1\},$$

则当 $i\in S$ 和 $j\in T$ 时我们有 $a_{ij}=0$, 于是 $\mathbf{A}$ 为可约的. 如果对某些 m, $a_{m,m+1}=0$, 便有类似的论断成立.

作为这个推论的一个自然结果,我们得到由 (7-14) 所定义的矩阵 $\mathbf{A}$ 与 (7-28) 所定义的 $\mathbf{A}(\mathbf{y})$ 为不可约的, 如果对 $n=1,\cdots,N-1$ 及 $\mu=0,2$, 分别有

$$1-h^2\beta_\mu g_n\neq 0 \text{ 和 } 1-h^2\beta_\mu f_y(x_n,y_n)\neq 0.$$

如果 L 表示函数 $f(x,y)$ 的一个 Lipschitz 常数, 这些条件当

$$h^2\beta_\mu L<1,\ \mu=0,2 \tag{7-32}$$

时是满足的。 当 $\beta_0=\beta_2=0$ 或当 $\beta_\mu\neq 0$ 时，对于充分小的 h 便是这种情形。

7.2-2. 单调矩阵. 我们回顾一下所用记号 $\mathbf{z}\geqslant 0$，我们指的是 $\mathbf{z}$ 的所有分量 z_i 均满足 $z_i\geqslant 0$. 类似地，记号 $\mathbf{z}>0$ 表示所有分量满足 $z_i>0$.

具有实的元素的矩阵 $\mathbf{A}$ 称为单调的，如果 $\mathbf{Az}\geqslant 0$ 蕴含着 $\mathbf{z}\geqslant 0$. 如 $\mathbf{A}$ 为单调的，且 $\mathbf{Az}\leqslant 0$，则由于

$$-\mathbf{Az}=\mathbf{A}(-\mathbf{z})\geqslant 0,$$

我们有 $-\mathbf{z}\geqslant 0$ 或 $\mathbf{z}\leqslant 0$. 因此，如 $\mathbf{A}$ 为单调的且 $\mathbf{Az}=0$，则 $\mathbf{z}$ 必定同时满足 $\mathbf{z}\geqslant 0$ 和 $\mathbf{z}\leqslant 0$，因此 $\mathbf{z}=0$. 于是

$$\det\mathbf{A}\neq 0,$$

我们便证明了：一个单调矩阵是非奇异的. 因此下面定理的陈述是有意义的.

定理 7.3. 一个矩阵 $\mathbf{A}$ 为单调当且仅当逆矩阵 $\mathbf{A}^{-1}$ 的元素为非负.

类似于上面的向量所使用的记号，定理的条件可以叙述为 $\mathbf{A}^{-1}\geqslant 0$.

证. (i) 令 $\mathbf{A}^{-1}\geqslant 0$, $\mathbf{Az}\geqslant 0$. 于是 $\mathbf{z}=\mathbf{A}^{-1}(\mathbf{Az})\geqslant$ ，因为一个非负矩阵作用于一个非负向量便产生一个非负向量. 由此推得 $\mathbf{A}$ 为单调的.

(ii) 假设 $\mathbf{A}^{-1}=(b_{ij})$ 有一个非负元素 b_{rs}，用 $\mathbf{e}_s$ 表示单位矩阵的第 s 列，则向量 $\mathbf{z}=\mathbf{A}^{-1}\mathbf{e}_s$ 的第 r 个元素等于 b_{rs}，因此为负. 于是 $\mathbf{Az}=\mathbf{A}(\mathbf{A}^{-1}\mathbf{e}_s)=\mathbf{e}_s$ 非负，但是 $\mathbf{z}$ 并不是非负的，所以 $\mathbf{A}$ 就不是单调的.

在大多数实际情形中，定理 7.3 所包含的准则不能用来决定给定的矩阵是否为单调的，由于逆矩阵 $\mathbf{A}^{-1}$ 并不是显式地知道的. 下面的结果包含了一个应用起来方便的 $\mathbf{A}$ 为单调的充分条件.

定理 7.4. 令矩阵 $\mathbf{A}=(a_{ij})$ 为不可约并满足条件：

(i) $a_{ij}\leqslant 0,\ i\neq j,\ i,j=1,\cdots,n,$

(ii) $\sum\limits_{j=1}^{n} a_{ij}\begin{cases}\geqslant 0, & i=1,2,\cdots,n,\\ >0, & \text{至少对某个 } i,\end{cases}$

则 $\mathbf{A}$ 为单调的.

条件 (ii) 意指 $\mathbf{A}$ 的每一行元素的和均为非负，并且至少有一行为正。在德国文献中，(ii) 称为“弱行和判则”.

作为定理 7.4 的一个结果，我们将可断言，由 (7-28) 所定义的矩阵 $\mathbf{A}(\mathbf{y})$ [这个矩阵包含 (7-14) 作为特殊情况]，对于 M 类边值问题是单调的. 对它来说，在前节中已经考虑到这些矩阵对于充分小的 h 为不可约的；对于同样的 h 值，非对角线上的元素为非正，其行和由

$$\sum_{j=1}^{N-1} a_{ij}=h^2\{\beta_0 f_y(x_{i-1},y_{i-1})+\beta_1 f_y(x_i,y_i)+\beta_2 f_y(x_{i+1},y_{i+1})\}$$

$$\text{当 } i=2,3,\cdots,N-2 \text{ 时} \tag{7-33}$$

以及

$$\sum_{j=1}^{N-1} a_{1,j}=1+h^2\{\beta_1 f_y(x_1,y_2)+\beta_2 f_y(x_2,y_2)\},$$

$$\sum_{j=1}^{N-1} a_{N-1,j}=1+h^2\{\beta_0 f_y(x_{N-1},y_{N-1})+\beta_1 f_y(x_N,y_N)\} \tag{7-34}$$

给出.

如果满足下面的条件 (7-40)，(7-33) 以及 (7-34) 右端的项分别为 $\geqslant 0$ 和 >0. 因此条件 (ii) 也被满足，由此得出 $\mathbf{A}(\mathbf{y})$ 为单调. 这个事实对 §7.3 中的误差估计具有决定性意义.

定理 7.4 的证明. 假设存在着一个向量 $\mathbf{z}$，具有非负分

量 $z_q < 0$, $q \in W$，使得 $\mathbf{Az} \geqslant 0$. 这个假设等价于假设 $\mathbf{A}$ 不是单调的；我们将证明这与 $\mathbf{A}$ 为不可约的假设相矛盾. 以 $\mathbf{e}$ 表示分量均为 1 的向量，由于 $\mathbf{Ae}$ 的分量正好是 $\mathbf{A}$ 的行和，故由 (ii) 就有 $\mathbf{Ae} \geqslant 0$, $\mathbf{Ae} \neq 0$. 由于两个非负向量的和为非负，由此得出，对于 $0 \leqslant \lambda \leqslant 1$，有

$$\lambda\mathbf{Az} + (1 - \lambda)\mathbf{Ae} = \mathbf{A}[\lambda\mathbf{z} + (1 - \lambda)\mathbf{e}] \geqslant 0. \tag{7-35}$$

把向量

$$\mathbf{w}_\lambda = \lambda\mathbf{z} + (1 - \lambda)\mathbf{e} \tag{7-36}$$

看作 λ 的函数. 当 $\lambda = 0$ 时 $\mathbf{w}_\lambda$ 的所有分量均为正，即为 $+1$；当 $\lambda = 1$ 时，则至少有一个负分量，即 z_q. $\mathbf{w}_\lambda$ 的分量为 λ 的连续函数. 当 λ 从 0 变化到 1 时，$\mathbf{w}_\lambda$ 至少有一个分量必定通过 0 值. 令 Λ 为使得 $\mathbf{w}_\lambda$ 有一个零分量的最小 λ 值. 显然 $0 < \Lambda < 1$. 现令 S 为 $\mathbf{w}_\lambda$ 的零分量下标的集合，并令 $\mathbf{T} = \mathbf{W} - \mathbf{S}$. 从它的构造来说，$\mathbf{S}' \neq \phi$. 而同时 $\mathbf{T} \neq \phi$，因为如果 $\mathbf{w}_\Lambda$ 的所有分量均为零，则向量 $\mathbf{z}$ 及 $\mathbf{e}$ 将会成比例，

$$\mathbf{e} = -\frac{\Lambda}{1 - \Lambda}\mathbf{z},$$

而从 $\mathbf{Az} \geqslant 0$ 将产生

$$\mathbf{Ae} = -\frac{\Lambda}{1 - \Lambda}\mathbf{Az} \leqslant \mathbf{0}.$$

与 (ii) 矛盾. 由 (7-35)，$\mathbf{Aw}_\Lambda \geqslant 0$；因此特别当 $i \in S$，

$$(\mathbf{Aw}_\Lambda)_i = \sum_{j \in T} a_{ij} w_{\Lambda j} \geqslant 0. \tag{7-37}$$

从其构造来说，当 $j \in T$ 时 $w_{\Lambda j} > 0$. 因此，(7-37) 由于 (i) 仅当 $a_{ij} = 0$, $i \in S$, $j = T$ 时才可能，亦即，如果 $\mathbf{A}$ 为可约的. 这一矛盾证明了定理的论断.

下面的结果将使我们当 $\mathbf{A}$ 满足定理 7.4 的假设时能够

得到 $\mathbf{A}^{-1}$ 元素的界.

定理 7.5. 令 $\mathbf{A}$ 和 $\mathbf{B}$ 为单调矩阵,并假设

$$\mathbf{A}-\mathbf{B}\geqslant 0, \tag{7-38}$$

那么

$$\mathbf{B}^{-1}-\mathbf{A}^{-1}\geqslant 0. \tag{7-39}$$

证. 由定理 7.3, 矩阵 $\mathbf{A}^{-1}$ 和 $\mathbf{B}^{-1}$ 存在且为非负. 非负矩阵的乘积是非负的. 从恒等式

$$\mathbf{B}^{-1}-\mathbf{A}^{-1}=\mathbf{B}^{-1}(\mathbf{A}-\mathbf{B})\mathbf{A}^{-1}$$

便得所需结果,由于右端的因子为非负.

回到我们的差分方程 (7-27), 规定

$$\beta_\mu\geqslant 0,\quad \mu=0,1,2. \tag{7-40}$$

如果矩阵 $\mathbf{A}(\mathbf{y})$ 由 (7-28) 确定,对于 M 类问题,我们有

$$\mathbf{A}(\mathbf{y})\geqslant \mathbf{J}. \tag{7-41}$$

如果 $h^2L\leqslant 1$, $\mathbf{A}(\mathbf{y})$ 满足定理 7.4 的假设,于是便得到

$$0\leqslant[\mathbf{A}(\mathbf{y})]^{-1}\leqslant \mathbf{J}^{-1}. \tag{7-42}$$

我们要求 $[\mathbf{A}(\mathbf{y})]^{-1}$ 的元素的界,因此将显式地确定 $\mathbf{J}^{-1}$. 以 $\mathbf{j}_m$ 表示 $\mathbf{J}^{-1}$ 的 m 列 ($m=1,2,\cdots,N-1$), 我们有

$$\mathbf{J}\mathbf{j}_m=\mathbf{e}_m, \tag{7-43}$$

其中 $\mathbf{e}_m$ 表示第 m 个单位向量. 将 (7-43) 用分量的形式写出,规定 $j_{0,m}=j_{N,m}=0$, 我们得到

$$-j_{n-1,m}+2j_{n,m}-j_{n+1,m}=\begin{cases}0, & n=1,\cdots,m-1,\\ 1, & n=m,\\ 0, & n=m+1,\cdots,N-1.\end{cases} \tag{7-44}$$

由此得 $j_{nm}=p_m n\ (n\leqslant m)$ 及 $j_{n,m}=q_m(N-n)\ (n\geqslant m)$,其中 p_m 和 q_m 为二个常数, 它们是这样确定的, 使得在 $n=m$ 时是相容的并且满足 (7-44) 中的一个条件. 这就导出条件

$$p_m m-q_m(N-m)=0,\quad p_m+q_m=1.$$

由此我们容易得到

$$p_m=\frac{N-m}{N},\ q_m=\frac{m}{N}.$$

因此我们有 $\mathbf{J}^{-1}=(j_{mn})$，其中

$$j_{mn}=\begin{cases}\dfrac{(N-m)n}{N}, & n\leqslant m,\\ \dfrac{m(N-n)}{N}, & n\geqslant m;\ m,n=1,2,\cdots,N-1.\end{cases}$$

(7-45)

记住 $n=(x_n-a)h^{-1}$ 和 $N-m=(b-x_m)h^{-1}$，我们可用下面的方法重述(7-42)：如果(7-2)和(7-40)成立，矩阵 $[\mathbf{A}(\mathbf{y})]^{-1}$ 的元素 d_{ij} 就满足

$$0\leqslant d_{ij}\leqslant\frac{(x_m-a)(b-x_n)}{h(b-a)},\tag{7-46}$$

其中 $m=\min(i,j)$ 和 $n=\max(i,j)$.

7.2-3. 一个新向量范数. 在讨论 Newton 方法的收敛性中，同时也为了以后要给出误差估计，有必要对具有分量为 $v_i(i=1,\cdots,n)$ 的向量 $\mathbf{v}$ 的大小进行估计. 在第三章中，为了这个目的，我们曾介绍了表达式

$$\|\mathbf{v}\|=|v_1|+|v_2|+\cdots+|v_n|,\tag{7-47}$$

然而对于所研究的问题，采用量

$$|\mathbf{v}|=\max_{1\leqslant i\leqslant n}|v_i|\tag{7-48}$$

来度量 $\mathbf{v}$ 的大小则更为方便. 用这个记号与一个数的绝对值的概念不会引起混淆，由于这个概念对向量来说并没有定义过. 显然关系式

$$|\mathbf{v}|\geqslant 0,\ |\mathbf{v}|=0 \text{ 当且仅当 } \mathbf{v}=0,\tag{7-49a}$$

$$|\mathbf{v}+\mathbf{w}|\leqslant|\mathbf{v}|+|\mathbf{w}|,\tag{7-49b}$$

$$|\lambda\mathbf{v}|=|\lambda||\mathbf{v}|\tag{7-49c}$$

是成立的，其中 $\mathbf{v}$ 和 $\mathbf{w}$ 是任意向量，而 λ 是任意一个实数或复数．在线性向量空间理论所使用的术语中，这些关系式刻划了数 $|\mathbf{v}|$ 为一个范数． 如果 $\mathbf{v}^{(\nu)}(\nu = 1, 2, \cdots)$ 是具有分量为 $v_i^{(\nu)}(i = 1, \cdots, n)$ 的向量序列，显然，当 $n \to \infty$ 时，

$$|\mathbf{v}^{(\nu)}| \to 0$$

当且仅当 $v_i^{(\nu)} \to 0(i = 1, \cdots, n)$．

如果 $\mathbf{A} = (a_{ij})$ 是具有实或复元素的一个矩阵，我们要求 C 是这样的一个数，它对所有向量 $\mathbf{v}$ 满足

$$|\mathbf{A}\mathbf{v}| \leqslant C|\mathbf{v}|. \tag{7-50}$$

我们断言，$C = |\mathbf{A}|$ 便具有这一性质，其中

$$|\mathbf{A}| = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{ij}|, \tag{7-51}$$

而且这是最好可能的 C 值(即最小的)．的确，

$$|\mathbf{A}\mathbf{v}| = \max_{1 \leqslant i \leqslant n} \left|\sum_{j=1}^{n} a_{ij}v_j\right| \leqslant \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{ij}||v_j|$$

$$\leqslant \max_{1 \leqslant i \leqslant n} \left(\max_{1 \leqslant j \leqslant n} |v_j| \sum_{j=1}^{n} |a_{ij}|\right) = |\mathbf{v}||\mathbf{A}|,$$

于是 (7-50) 便被满足．另一方面，如果 $\mathbf{A} = \mathbf{0}$，$C = |\mathbf{A}|$ 显然是最好可能的值；如果 $\mathbf{A} \neq \mathbf{0}$，令 i 是使得

$$|\mathbf{A}| = \sum_{j=1}^{n} |a_{ij}|$$

成立，并定义 $\mathbf{v}$ 为

$$v_j = 0, \text{如果 } a_{ij} = 0; \quad v_j = \frac{\bar{a}_{ij}}{|a_{ij}|}, \text{如果 } a_{ij} \neq 0,$$

于是 $|\mathbf{v}| = 1$，并且 $\mathbf{A}\mathbf{v}$ 的第 i 个分量由

$$\sum_{\substack{j=1 \\ a_{ij} \neq 0}}^{n} a_{ij} \frac{\bar{a}_{ij}}{|a_{ij}|} = \sum_{j=1}^{n} |a_{ij}| = |\mathbf{A}|$$

给出．因此，对于这个特殊向量 $\mathbf{v}$，$|\mathbf{A}\mathbf{v}| \geqslant |\mathbf{A}||\mathbf{v}|$，(7-50)

不可能对任何比$|\mathbf{A}|$还小的C成立。

除关系式

$$|\mathbf{A}| \geqslant 0,\ |\mathbf{A}| = 0 \text{ 当且仅当 } \mathbf{A} = 0, \tag{7-52a}$$

$$|\mathbf{A} + \mathbf{B}| \leqslant |\mathbf{A}| + |\mathbf{B}|, \tag{7-52b}$$

$$|\lambda \mathbf{A}| = |\lambda| |\mathbf{A}| \tag{7-52c}$$

之外，它们类似于(7-49)并可作类似的证明. 对于矩阵$\mathbf{A}$，用(7-51)所联结的数$|\mathbf{A}|$还具有性质

$$|\mathbf{AB}| \leqslant |\mathbf{A}||\mathbf{B}|. \tag{7-53}$$

利用(7-50)和(7-51)，注意到

$$|\mathbf{A}| = \max_{|\mathbf{y}|=1} |\mathbf{Ay}|$$

就可以证明这一点. 因此有

$$|\mathbf{AB}| = \max_{|\mathbf{y}|=1} |\mathbf{ABy}| \leqslant |\mathbf{A}| \max_{|\mathbf{y}|=1} |\mathbf{By}| = |\mathbf{A}||\mathbf{B}|,$$

这就得到了证明.

关系式(7-52)和(7-53)阐明，用线性向量空间的术语，数$|\mathbf{A}|$便是线性算子$\mathbf{A}$的一个范数. 由于不会用到算子的其它范数，我们简称$|\mathbf{A}|$为矩阵$\mathbf{A}$的范数.

将要用到下面的引理(Banach 的一个非常一般结果的一个特殊情形).

引理 7.1. 令$\mathbf{A}$为一个矩阵，且满足$|\mathbf{A}| = k < 1$，并令$\mathbf{I}$表示单位矩阵，则存在矩阵$(\mathbf{I} - \mathbf{A})^{-1}$，并有

$$|(\mathbf{I} - \mathbf{A})^{-1}| \leqslant \frac{1}{1-k}. \tag{7-54}$$

证. 令

$$\mathbf{S}_p = \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^{p-1},$$

则

$$\mathbf{S} = \lim_{p \to \infty} \mathbf{S}_p$$

存在，由于

$$|\mathbf{S}_{p+q}-\mathbf{S}_p| = |\mathbf{A}^p+\cdots+\mathbf{A}^{p+q-1}|$$
$$\leqslant |\mathbf{A}^p|+\cdots+|\mathbf{A}^{p+q-1}| \leqslant \frac{k^p}{1-k}, \qquad (7\text{-}55)$$

在恒等式

$$(\mathbf{I}-\mathbf{A})\mathbf{S}_p = \mathbf{I}-\mathbf{A}^p$$

中，令 $p\to\infty$ 便得到

$$(\mathbf{I}-\mathbf{A})\mathbf{S} = \mathbf{I}.$$

因此矩阵 $\mathbf{S}$ 是右逆的，于是它就是矩阵 $\mathbf{I}-\mathbf{A}$ 的逆. 在(7-55)中令 $p=0$ 及 $q\to\infty$，便得到关系式(7-54).

我们需要一个具有略为特殊一些性质的进一步结果. 如果矩阵 $\mathbf{A}(\mathbf{y})$ 的元素 a_{ij} 在区域 B 中是 $\mathbf{y}$ 的连续可微函数，而 B 包含所有向量 $t\mathbf{y}_1+(1-t)\mathbf{y}_2$，其中 $\mathbf{y}_1$ 和 $\mathbf{y}_2$ 是给定的二个向量，而 $0\leqslant t\leqslant 1$，则有

$$|\mathbf{A}(\mathbf{y}_2)-\mathbf{A}(\mathbf{y}_2)| \leqslant K|\mathbf{y}_2-\mathbf{y}_1|, \qquad (7\text{-}56)$$

其中

$$K = \max_{\substack{1\leqslant i\leqslant n\\ \mathbf{y}\in B}} \sum_{j=1}^{n}\sum_{k=1}^{n}\left|\frac{\partial a_{ij}}{\partial y_k}\right|. \qquad (7\text{-}57)$$

对 $\mathbf{A}(\mathbf{y}_2)-\mathbf{A}(\mathbf{y}_1)$ 的每个元素应用中值定理，便可完成这个证明.

7.2-4. 非线性方程组的 Newton 方法. 假设关于 n 个未知量 $y_1,\cdots,y_n$ 的 n 个方程

$$\varphi_i(y_1,y_2,\cdots,y_n)=0,\ i=1,\cdots,n$$

可以写成向量形式

$$\boldsymbol{\varphi}(\mathbf{y}) = \mathbf{0}. \qquad (7\text{-}58)$$

令 $\mathbf{A}(\mathbf{y})=(a_{ij})$ 表示具有元素为

$$a_{ij} = \frac{\partial\varphi_i(\mathbf{y})}{\partial y_j}$$

的矩阵.

如果把向量 $\mathbf{y}=\mathbf{y}^{(0)}$ 看成是方程组 (7-58) 的一个近似解，并且矩阵 $\mathbf{A}(\mathbf{y}^{(0)})$ 是非奇异的，那么可希望通过对方程组 (7-58) 在 $\mathbf{y}=\mathbf{y}^{(0)}$ 处线性化得到的向量

$$\mathbf{y}^{(1)}=\mathbf{y}^{(0)}-[\mathbf{A}(\mathbf{y}^{(0)})]^{-1}\boldsymbol{\varphi}(\mathbf{y}^{(0)}) \tag{7-59}$$

为解的一个较好的近似．如果与之有关的矩阵 $\mathbf{A}(\mathbf{y}^{(\nu)})$ 保持非奇异，便可希望得到一个更好的近似值序列

$$\mathbf{y}^{(\nu)}(\nu=1,2,\cdots),$$

通过算法

$$\mathbf{y}^{(\nu+1)}=\mathbf{y}^{(\nu)}-[\mathbf{A}(\mathbf{y}^{(\nu)})]^{-1}\boldsymbol{\varphi}(\mathbf{y}^{(\nu)}),\ \nu=0,1,2,\cdots. \tag{7-60}$$

它被称为解非线性方程组 (7-58) 的 Newton 方法．关于 Newton 方法收敛于方程组 (7-58) 的一个解的问题，要得到一个简单的充分条件，在 L. V. Kantorovich 1937 年发表一个非常一般的情形下，甚至无需假设解的存在的情形下，保证 Newton 方法收敛定理(见 Kantorovich[1948]) 之前，一直都被认为是数值分析的一个困难问题．我们将以适合于当前问题的形式来叙述 Kantorovich 的结果，这就是当

$$\boldsymbol{\varphi}(\mathbf{y})=\mathbf{r}(\mathbf{y})$$

时讨论 Newton 方法的收敛性，其中 $\mathbf{r}(\mathbf{y})$ 是由 (7-25) 定义的．

定理 7.6. 假设满足下面的条件：

(i) 对于初始近似值 $\mathbf{y}=\mathbf{y}^{(0)}$，矩阵 $\mathbf{A}(\mathbf{y}^{(0)})$ 有一个逆

$$\boldsymbol{\Gamma}_0=\mathbf{A}(\mathbf{y}^{(0)})^{-1},$$

并且对其范数的估计已知为

$$|\boldsymbol{\Gamma}_0|\leqslant B_0; \tag{7-61}$$

(ii) 向量 $\mathbf{y}^{(0)}$ 在

$$|\boldsymbol{\Gamma}_0\boldsymbol{\varphi}(\mathbf{y}^{(0)})|\leqslant\eta_0 \tag{7-62}$$

的意义下近似地满足方程组 (7-58)；

(iii) 在如下不等式 (7-65) 所定义的区域中，向量 $\boldsymbol{\varphi}(\mathbf{y})$ 的分量对 $\mathbf{y}$ 是二次连续可微的，并满足

$$\sum_{j,k=1}^{n}\left|\frac{\partial^2\varphi_i}{\partial y_j\partial y_k}\right|\leqslant K,\ i=1,2,\cdots,n; \qquad (7\text{-}63)$$

(iv) 上面所引进的常数 B_0, η_0 及 K 满足不等式

$$h_0\equiv B_0\eta_0K\leqslant\frac{1}{2}, \qquad (7\text{-}64)$$

那么方程组 (7-58) 有一个解 $\mathbf{y}^*$，它位于立方体

$$|\mathbf{y}-\mathbf{y}^{(0)}|\leqslant N(h_0)\eta_0=\frac{1-\sqrt{1-2h_0}}{h_0}\eta_0 \qquad (7\text{-}65)$$

中。此外，由 (7-60) 所定义的逐次近似值 $\mathbf{y}^{(\nu)}$ 存在且收敛于 $\mathbf{y}$，而它的收敛速度可用不等式

$$|\mathbf{y}^{(\nu)}-\mathbf{y}^*|\leqslant\frac{1}{2^{\nu-1}}(2h_0)^{2^\nu-1}\eta_0 \qquad (7\text{-}66)$$

来估计。甚至在一维情形 ($n=1$)，这个定理也不是明显的。

证。令

$$B_1=\frac{B_0}{1-h_0},\ \eta_1=\frac{1}{2}\frac{h_0\eta_0}{1-h_0},\ h_1=\frac{1}{2}\frac{h_0^2}{(1-h_0)^2}. \qquad (7\text{-}67)$$

我们将首先证明，如果将每处的下标 0 都换成 1，则关系式 (7-61), (7-62) 及 (7-64) 仍成立。

首先由 (7-59) 有

$$|\mathbf{y}^{(1)}-\mathbf{y}^{(0)}|=|\boldsymbol{\Gamma}_0\boldsymbol{\varphi}(\mathbf{y}^{(0)})|\leqslant\eta_0. \qquad (7\text{-}68)$$

此外，在目前的意义下，对于 $\mathbf{A}(\mathbf{y})$ 利用 (7-56)，

$$|\boldsymbol{\Gamma}_0[\mathbf{A}(\mathbf{y}^{(0)})-\mathbf{A}(\mathbf{y}^{(1)})]|\leqslant|\boldsymbol{\Gamma}_0||\mathbf{A}(\mathbf{y}^{(0)})-\mathbf{A}(\mathbf{y}^{(1)})|$$
$$\leqslant B_0K|\mathbf{y}^{(0)}-\mathbf{y}^{(1)}|\leqslant B_0K\eta_0=h_0<1.$$

由 Banach 的引理 7.1 (§7.2-3)，得到矩阵

$$\mathbf{H}=\mathbf{I}-\boldsymbol{\Gamma}_0[\mathbf{A}(\mathbf{y}^{(0)})-\mathbf{A}(\mathbf{y}^{(1)})]$$

有一个逆，且满足

$$|\mathbf{H}^{-1}| \leqslant \frac{1}{1-h_0}. \tag{7-69}$$

令 $\boldsymbol{\Gamma}_1 = \mathbf{H}^{-1}\boldsymbol{\Gamma}_0$，由于 $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$，我们得到

$$\begin{aligned}
\boldsymbol{\Gamma}_1 &= [\mathbf{I} - \boldsymbol{\Gamma}_0(\mathbf{A}(\mathbf{y}^{(0)}) - \mathbf{A}(\mathbf{y}^{(1)}))]^{-1}\mathbf{A}(\mathbf{y}^{(0)})^{-1} \\
&= \{\mathbf{A}(\mathbf{y}^{(0)})[\mathbf{I} - \boldsymbol{\Gamma}_0(\mathbf{A}(\mathbf{y}^{(0)}) - \mathbf{A}(\mathbf{y}^{(1)}))]\}^{-1} \\
&= \{\mathbf{A}(\mathbf{y}^{(0)}) - \mathbf{A}(\mathbf{y}^{(0)}) + \mathbf{A}(\mathbf{y}^{(1)})\}^{-1} \\
&= [\mathbf{A}(\mathbf{y}^{(1)})]^{-1}.
\end{aligned}$$

因而证明了矩阵 $\mathbf{A}(\mathbf{y}^{(1)})$ 存在一个逆矩阵 $\boldsymbol{\Gamma}_1 = \mathbf{H}^{-1}\boldsymbol{\Gamma}_0$。利用(7-69)，便得到

$$|[\mathbf{A}(\mathbf{y}^{(1)})]^{-1}| = |\mathbf{H}^{-1}\boldsymbol{\Gamma}_0| \leqslant \frac{B_0}{1-h_0} = B_1.$$

所以(7-61)在下标升高1时成立。

将 $\boldsymbol{\varphi}(\mathbf{y})$ 的每个分量在 $\mathbf{y} = \mathbf{y}^{(0)}$ 处按 Taylor 公式展开，便得到

$$\boldsymbol{\varphi}(\mathbf{y}) = \boldsymbol{\varphi}(\mathbf{y}^{(0)}) + \mathbf{A}(\mathbf{y}^{(0)})(\mathbf{y} - \mathbf{y}^{(0)}) + \frac{1}{2}\mathbf{r}(\mathbf{y}), \tag{7-70}$$

其中 $\mathbf{r}$ 的第 i 个分量为

$$r_i = \sum_{j,k=1}^{n} \frac{\partial^2 \varphi_i(\mathbf{y}^{(0)} + \theta_i(\mathbf{y} - \mathbf{y}^{(0)}))}{\partial y_j \partial y_k}(y_j - y_j^{(0)})(y_k - y_k^{(0)}),$$

$$0 < \theta_i < 1.$$

因此，如果 $\mathbf{y}$ 满足(7-65)，就有

$$|\mathbf{r}| \leqslant K|\mathbf{y} - \mathbf{y}^{(0)}|^2.$$

由于 $\boldsymbol{\varphi}(\mathbf{y}^{(0)}) + \mathbf{A}(\mathbf{y}^{(0)})(\mathbf{y}^{(1)} - \mathbf{y}^{(0)}) = \mathbf{0}$，得到

$$\boldsymbol{\varphi}(\mathbf{y}^{(1)}) = \frac{1}{2}\mathbf{r}(\mathbf{y}^{(1)}).$$

于是推出结果

$$|\boldsymbol{\Gamma}_0\boldsymbol{\varphi}(\mathbf{y}^{(1)})| \leqslant \frac{1}{2} B_0 K \eta_0^2 = \frac{1}{2} h_0\eta_0.$$

因此，最后由于 $h_0 \leqslant \frac{1}{2}$，

$$|\boldsymbol{\Gamma}_1\boldsymbol{\varphi}(\mathbf{y}^{(1)})| \leqslant |\mathbf{H}^{-1}\boldsymbol{\Gamma}_0\boldsymbol{\varphi}(\mathbf{y}^{(1)})| \leqslant |\mathbf{H}^{-1}||\boldsymbol{\Gamma}_0\boldsymbol{\varphi}(\mathbf{y}^{(0)})|$$
$$\leqslant \frac{1}{2}\frac{1}{1-h_0} h_0\eta_0 = \eta_1 < \eta_0.$$

于是便证明了(7-62)当下标升高 1 时是满足的。

如果 $\mathbf{y}$ 位于由

$$|\mathbf{y}-\mathbf{y}^{(1)}| \leqslant \frac{1-\sqrt{1-2h_1}}{h_1}\eta_1$$

定义的区域中，其中 h_1 由(7-67)确定，那么通过使用(7-68)以及 h_1 和 η_1 的定义，得到

$$|\mathbf{y}-\mathbf{y}^{(0)}| \leqslant |\mathbf{y}-\mathbf{y}^{(1)}| + |\mathbf{y}^{(1)}-\mathbf{y}^{(0)}|$$
$$\leqslant \frac{1-\sqrt{1-2h_1}}{h_1}\eta_1 + \eta_0 = \frac{1-\sqrt{1-2h_0}}{h_0}\eta_0.$$

因而 $\mathbf{y}$ 仍位于由(7-65)定义的区域中．因此，(7-65)当下标上升 1 时仍成立．最后我们也有

$$h_1 = B_1\eta_1 K = \frac{B_0}{1-h_0}\frac{1}{2}\frac{h_0\eta_0}{1-h_0}K$$
$$= \frac{1}{2}\frac{h_0^2}{(1-h_0)^2} \leqslant 2h_0^2 \leqslant \frac{1}{2}.$$

故(7-64)也满足．

用 $\mathbf{y}^{(1)}$，$\mathbf{y}^{(2)}$ 来代替 $\mathbf{y}^{(0)}$，$\mathbf{y}^{(1)}$，显然可以重复上述的论证．如果数 η_ν 和 h_ν 由

$$\eta_\nu = \frac{1}{2}\frac{h_{\nu-1}\eta_{\nu-1}}{1-h_{\nu-1}},\ h_\nu = \frac{1}{2}\frac{h_{\nu-1}^2}{(1-h_{\nu-1})^2},\ \nu = 1,2,\cdots$$

递推地确定，我们便得到

$$|\mathbf{y}^{(\nu+1)}-\mathbf{y}^{(\nu)}| \leqslant \eta_\nu,\ \nu = 1,2,\cdots.$$

利用归纳法，容易证明

$$h_\nu \leqslant \frac{1}{2}(2h_0)^{2^\nu}, \quad \nu = 1, 2, \cdots.$$

因此

$$\frac{1}{2}\frac{1}{1-h_{\nu-1}} \leqslant 1,$$

并且

$$\begin{aligned}\eta_\nu &\leqslant h_{\nu-1}\eta_{\nu-1} \leqslant \cdots \leqslant h_{\nu-1}\cdots h_0\eta_0 \\ &\leqslant \frac{1}{2^\nu}(2h_0)^{2^{\nu-1}+2^{\nu-2}+\cdots+1}\eta_0 = \frac{1}{2^\nu}(2h_0)^{2^\nu-1}\eta_0.\end{aligned}$$

由此得出对任何整数 $p \geqslant 1$，有

$$\begin{aligned}|\mathbf{y}^{(\nu+p)} - \mathbf{y}^{(\nu)}| &\leqslant |\mathbf{y}^{(\nu+p)} - \mathbf{y}^{(\nu+p-1)}| \\ &\quad + \cdots + |\mathbf{y}^{(\nu+1)} - \mathbf{y}^{(\nu)}| \\ &\leqslant \eta_{\nu+p-1} + \eta_{\nu+p-2} + \cdots + \eta_\nu.\end{aligned}$$

使用代数恒等式

$$\begin{aligned}&\eta_\nu\frac{1-\sqrt{1-2h_\nu}}{h_\nu} - \eta_{\nu+1}\frac{1-\sqrt{1-2h_{\nu+1}}}{h_{\nu+1}} \\ &= \eta_\nu, \quad \nu = 1, 2, \cdots,\end{aligned}$$

便可导出

$$\begin{aligned}|\mathbf{y}^{(\nu+p)} - \mathbf{y}^{(\nu)}| &\leqslant \eta_\nu\frac{1-\sqrt{1-2h_\nu}}{h_\nu} - \eta_{\nu+p}\frac{1-\sqrt{1-2h_{\nu+p}}}{h_{\nu+p}} \\ &\leqslant \eta_\nu\frac{1-\sqrt{1-2h_\nu}}{h_\nu}.\end{aligned}$$

或使用

$$\frac{1-\sqrt{1-2h_\nu}}{h_\nu} \leqslant 2,$$

得到

$$|\mathbf{y}^{(\nu+p)} - \mathbf{y}^{(\nu)}| \leqslant 2\eta_\nu \leqslant \frac{1}{2^{\nu-1}}(2h_0)^{2^\nu-1}\eta_0. \qquad (7\text{-}71)$$

由于右边的项与 p 无关且当 $\nu\to\infty$ 时趋向零，由此便得到以 $\mathbf{y}^{(\nu)}$ 的分量所形成的 n 个序列 $\{y_i^{(\nu)}\}$ 为 Cauchy 序列，因而有极限 $y_i^*(i=1,2,\cdots,n)$. 记以 y_i^* 为分量的向量为 $\mathbf{y}^*$，因此我们得到

$$\lim_{\nu\to\infty}\mathbf{y}^{(\nu)}=\mathbf{y}^*.$$

由于所有的 $\mathbf{y}^{(\nu)}$ 属于由 (7-65) 定义的紧致集，$\mathbf{y}^*$ 也属于它. 在 (7-71) 中令 $p\to\infty$ 便得到不等式 (7-66).

剩下的是要证明 $\mathbf{y}^*$ 是给定方程 (7-58) 的一个解. 考虑到恒等式

$$\boldsymbol{\varphi}(\mathbf{y}^{(\nu)})=\mathbf{A}(\mathbf{y}^{(\nu)})(\mathbf{y}^{(\nu)}-\mathbf{y}^{(\nu+1)}),$$

便可完成这个证明. 使用 (7-56)，我们有

$$\begin{aligned}|\mathbf{A}(\mathbf{y}^{(\nu)})|&\leqslant|\mathbf{A}(\mathbf{y}^{(0)})|+|\mathbf{A}(\mathbf{y}^{(\nu)})-\mathbf{A}(\mathbf{y}^{(0)})|\\&\leqslant|\mathbf{A}(\mathbf{y}^{(0)})|+K\frac{1-\sqrt{1-2h_0}}{h_0}\eta_0=C.\end{aligned}$$

因此

$$|\boldsymbol{\varphi}(\mathbf{y}^{(\nu)})|\leqslant C|\mathbf{y}^{(\nu)}-\mathbf{y}^{(\nu+1)}|.$$

令 $\nu\to\infty$ 且利用 $\boldsymbol{\varphi}(\mathbf{y})$ 是连续的这个事实，我们得到

$$|\boldsymbol{\varphi}(\mathbf{y}^*)|=0,\ \text{因而}\ \boldsymbol{\varphi}(\mathbf{y}^*)=0.$$

定理 7.6 证毕.

如果由 (7-66) 给出关于 $|\mathbf{y}^{(\nu)}-\mathbf{y}^*|$ 的界记成 b_ν，便有

$$\log b_\nu=(\nu-1)\log\frac{1}{2}+(2^\nu-1)\log 2h_0+\log\eta_0.$$

因此，如果 $2h_0<1$，那么

$$\frac{\log b_{\nu+1}}{\log b_\nu}\to 2,\ \nu\to\infty. \qquad (7-72)$$

一般地，如果 b_ν 是在解方程的一个迭代方法的第 ν 步上的误差，极限 (7-72) 的最好可能达到的值称为方法收敛的阶. 考虑到简单的例子(见问题 14)，易证对 Newton 方法来

说，(7-72) 中的值 2 是无法改进的，因此 Newton 方法的收敛阶为 2。这个事实有时可表示成方法为平方地收敛的或是"在每一步上正确的小数位数加倍"。

可以证明，如果在 Newton (7-60) 中，矩阵 $\mathbf{A}(\mathbf{y}^{(\nu)})$ 在每一步上换成 $\mathbf{A}(\mathbf{y}^{(0)})$ 所得序列 $\mathbf{y}^{(1)},\mathbf{y}^{(2)},\cdots$ 仍然收敛于 $\mathbf{y}^*$，假设初始近似值 $\mathbf{y}^{(0)}$ 充分接近的话。然而，这个变形的 Newton 过程的收敛阶仅为 1 (见问题 16)。

7.2-5. 关于差分方程组 (7-24) 的 Newton 方法的收敛性. 我们将证明，在某些假设下，定理 7.6 的条件对 (非线性) 差分方程组 (7-24) 是满足的。我们将始终假设 (7-40) 是成立的，这样 $\mathbf{A}(\mathbf{y})$ 便是单调的。考虑到 (7-42)，由于

$$\sum_{n=1}^{N-1} j_{mn} = \frac{m(N-m)}{2} \leqslant \frac{N^2}{8},\quad m=1,2,\cdots,N-1,$$

便得到 (7-61) 当

$$B_0 = \frac{(b-a)^2}{8h^2} \tag{7-73}$$

时成立。由于

$$\begin{aligned}\varphi_n(\mathbf{y}) = -y_{n-1} + 2y_n - y_{n+1} + h^2\{\beta_0 f(x_{n-1},y_{n-1})\\ + \beta_1 f(x_n,y_n) + \beta_2 f(x_{n+1},y_{n+1})\},\end{aligned}$$

又考虑到 $\beta_0+\beta_1+\beta_2=1$，条件 (7-63) 为

$$K = h^2 L_2 = h^2 \max_{\substack{x\in[a,b]\\ -\infty<y<\infty}} |f_{yy}(x,y)| \tag{7-74}$$

所满足。令初始近似值 $\mathbf{y}^{(0)}$ 由

$$y_n^{(0)} = z(x_n),\quad n=1,\cdots,N-1 \tag{7-75}$$

确定，其中 $z(x)$ 为一个 $p+2$ 次可微函数，且满足 $z(a)=A$，$z(b)=B$，p 为差分算子的阶，并令

$$\begin{aligned} Z &= \max_{a\leqslant x\leqslant b} |z^{(p+2)}(x)|,\\ R &= \max_{a\leqslant x\leqslant b} |z''(x) - f(x,z(x))|,\end{aligned} \tag{7-76}$$

因而由 §6.1 的结果推得，对于某个常数 G，有

$$|1 - z(x-h) + 2z(x) - z(x+h) + h^2\{\beta_0 z''(x-h) + \beta_1 z''(x) + \beta_2 z''(x+h)\}| \leqslant h^{p+2}GZ.$$

考虑到 $\beta_i \geqslant 0$，$\beta_0 + \beta_1 + \beta_2 = 1$，于是

$$\begin{aligned} |r_n(y^{(0)})| = | &- z(x_{n-1}) + 2z(x_n) - z(x_{n+1}) \\ &+ h^2\{\beta_0 f(x_{n-1}, z(x_{n-1})) + \beta_1 f(x_n, z(x_n)) \\ &+ \beta_2 f(x_{n+1}, z(x_{n+1}))\}| \\ \leqslant\ & h^2 R + h^{p+2}GZ,\quad n = 1,2,\cdots,N-1. \end{aligned}$$

利用 (7-73)，当

$$\eta_0 = \frac{(b-a)^2}{8}(R + h^p GZ) \tag{7-77}$$

时，(7-62) 便被满足．于是保证 Newton 方法收敛的 Kantorovich 定理的主要条件 (7-64)，当

$$\frac{(b-a)^4 L_2(R + h^p GZ)}{64} \leqslant \frac{1}{2} \tag{7-78}$$

时便得到满足。

从这个不等式中可以引出各种结论，如果选取次数 $< (p+2)$ 的多项式 $z(x)$ 作为初始近似值，则 $Z = 0$．于是这个方法至少是收敛的，如果 R 这个量(它可以解释为使微分方程不满足时误差的最大值)满足

$$R \leqslant \frac{32}{(b-a)^4 L_2}. \tag{7-79}$$

为了理论的需要，我们可以假设初始近似 $z(x)$ 为精确解 $y(x)$．因而我们得到 $R = 0$．由 (7-78)，当

$$h^p \leqslant \frac{32}{(b-a)^4 L_2 GZ} \tag{7-80}$$

时，即对所有充分小的 h 值，Newton 方法收敛于差分方程组的一个解。

于是有限差分格式的解的存在性已经证明完毕．现在我们将证明M类问题的这个解的唯一性．事实上，如果 $\mathbf{y}$ 和 $\mathbf{z}$ 为任何两个解；那么便可令

$$f(x_n, z_n) - f(x_n, y_n) = \eta_n(z_n - y_n),$$

其中 η_n 为 f_y 的一个值．利用(7-2)和 Lipschitz 条件，

$$0 \leqslant \eta_n \leqslant L,$$

因而易见向量 $\mathbf{d} = \mathbf{z} - \mathbf{y}$ 满足方程组

$$(\mathbf{J} + h^2\mathbf{BH})\mathbf{d} = 0,$$

其中 $\mathbf{J}$ 和 $\mathbf{B}$ 是在 §7.1-3 中定义的，而 $\mathbf{H}$ 是以 η_n 为对角线元素的对角矩阵． 由定理 7.4，矩阵 $\mathbf{J} + h^2\mathbf{BH}$ 当 $h^2 \leqslant L^{-1}$ 时为单调的，因而是非奇异的．由此得出 $\mathbf{d} = \mathbf{0}$．

我们将这些陈述概括在下面的定理中，这是本节的主要结果．

定理 7.7. 设有限差分方程组(7-24)已由M类的一个边值问题所导出，若 p 表示有限差分算子(7-8)的阶，假设精确解 $y(x)$在 $[a,b]$ 内有一个连续的 $(p+2)$ 阶导数，且令

$$Z = \max_{a \leqslant x \leqslant b} |y^{(p+2)}(x)|, \tag{7-81}$$

假设G由 §6.1-3 中定义，且有

$$\beta_i \geqslant 0,\ i = 0,1,2 \ \text{和}\ Lh^2 < 1, \tag{7-82}$$

其中L表示 $f(x,y)$ 的 Lipschitz 常数． 若 L_2 为 $f_{yy}(x, y)$ 在 $a \leqslant x \leqslant b$, $-\infty < y < \infty$ 中的一个上界，则(7-24)对所有满足(7-80)的 h 存在唯一的解． 假设初始近似值由(7-75)确定，其中 $z(x)$满足(7-78)，这个解可通过 Newton 方法来得到．

这个定理包含 $L_2 = 0$ 这个特殊情形下线性边值问题唯一解的存在性．于是 Newton 方法取一步便可收敛．

7.3. M类边值问题的离散误差

如前所述，我们所说离散误差是指量 $e_n = y_n - y(x_n)$，其中 y_n 为由有限差分格式 (7-24) 所定的精确解（计算时没有舍入），而 $y(x)$ 为边值问题的精确解．我们将确定离散误差的界并建立它的渐近公式．

7.3·1．一个先验界． 离散误差的一个粗糙的界可以作为 Newton 方法收敛(定理 7.6)的一个推论得出．如果 Newton 方法从 $z(x) = y(x)$出发,用 $\mathbf{e}$ 表示分量为 e_n 的向量,则有

$$|\mathbf{y}^{(0)} - \mathbf{y}^*| = |\mathbf{e}|.$$

假设 (7-80) 成立,由 (7-65) 并利用 (7-77) 便得到

$$|\mathbf{e}| \leqslant \frac{1-\sqrt{1-2h_0}}{h_0} \frac{(b-a)^2 GZh^p}{8}, \qquad (7-83)$$

其中 h_0 由 (7-64) 定义．利用

$$1 < \frac{1-\sqrt{1-2h_0}}{h_0} \leqslant 2, \text{ 当 } 0 < h_0 \leqslant \frac{1}{2} \text{ 时},$$

从 (7-83) 中能导出简单的界

$$|\mathbf{e}| \leqslant \frac{(b-a)^2 GZh^p}{4}. \qquad (7-84)$$

现在我们将在无需使用 Newton 方法的路子来证明这个略微改进的且具有一般化形式的界．

定理 7.8． 不用 (7-24)，而令 y_n 值满足方程

$$-y_{n-1} + 2y_n - y_{n+1} + h^2\{\beta_0 f(x_{n-1}, y_{n-1}) + \beta_1 f(x_n, y_n) + \beta_2 f(x_{n+1}, y_{n+1})\} = \theta_n K h^{q+2}, \quad n = 0, 1, \cdots, N-1, \qquad (7-85)$$

其中 θ_n 为满足 $|\theta_n| \leqslant 1$ 的任意数，K 和 q 是任意非负常数．那么采用定理 7.7 中使用的记号，如果 (7-80) 成立，则离散

误差满足

$$|e_n| \leqslant \frac{(x_n - a)(b - x_n)}{2}(GZh^p + Kh^q),$$

$$n = 1, 2, \cdots, N - 1. \tag{7-86}$$

证．按照 §6.1-3，精确解 $y(x)$ 满足

$$\begin{aligned} -y(x_{n-1}) + 2y(x_n) - y(x_{n+1}) + h^2\{\beta_0 f(x_{n-1}, y(x_{n-1})) \\ + \beta_1 f(x_n, y(x_n)) + \beta_2 f(x_{n+1}, y(x_{n+1}))\} \\ = \theta'_n GZh^{p+2}, \end{aligned} \tag{7-87}$$

其中 $|\theta'_n| \leqslant 1$． 从 (7-85) 减去 (7-87)，我们对差

$$f(x_n, y_n) - f(x_n, y(x_n))$$

应用中值定理得到

$$\begin{aligned} -e_{n-1} + 2e_n + e_{n+1} + h^2\{\beta_0 g_{n-1} e_{n-1} + \beta_1 g_n e_n + \beta_2 g_{n+1} e_{n+1}\} \\ = \theta''_n (h^{q+2}K + h^{p+2}GZ), \end{aligned}$$

其中 $0 \leqslant g_n \leqslant L$． 用 $\mathbf{G}$ 表示具有元素为

$$g_n (n = 1, \cdots, N - 1)$$

的对角矩阵，并象 §7.1-3 一样来定义矩阵 $\mathbf{J}$ 和 $\mathbf{B}$，因此便得到

$$(\mathbf{J} + h^2\mathbf{BG})\mathbf{e} = (h^{q+2}K + h^{p+2}GZ)\boldsymbol{\theta},$$

其中 $\boldsymbol{\theta}$ 为具有分量数值不超过 1 的向量． 当 $h^2L \leqslant 1$ 时，矩阵 $\mathbf{J} + h^2\mathbf{BG}$ 满足定理 7.4 的条件，此外我们还有

$$\mathbf{J} + h^2\mathbf{BG} \geqslant \mathbf{J}.$$

由定理 7.5，得到

$$0 \leqslant (\mathbf{J} + h^2\mathbf{BG})^{-1} \leqslant \mathbf{J}^{-1}.$$

若 $\mathbf{J}^{-1} = (j_{mn})$，我们有

$$|e_m| \leqslant (h^{q+2}K + h^{p+2}GZ)\sum_{n=1}^{N-1} j_{mn}.$$

现在由

$$\sum_{n=1}^{N-1} j_{mn} = \frac{1}{N}[(N-m)(1+2+\cdots+m) + m(1+2+\cdots+N-m-1)] = \frac{m(N-m)}{2} = \frac{(x_m-a)(b-x_m)}{2h^2}$$

得出定理 7.8 的论断。

当 $K=0$ 时（差分方程被精确地满足），对于小的 h，以及 $x_m = \frac{1}{2}(a+b)$，界 (7-86) 实际上等价于 (7-84)。请读者证明对于边值问题 $y'' = x^{p+2}$，$y(a)=0$，$y(b)=0$，这个界是明显的。

7.3-2. 离散误差的渐近性态. 除去定理 7.8 的假设外，现在我们还假设阶 $p \geqslant 2$ 和 $K=0$，于是差分方程 (7-24) 精确地满足. 由此便得 $|\mathbf{e}| = O(h^2)$. 同时我们将假设差分算子满足

$$\beta_0 = \beta_2 \tag{7-88}$$

以及问题的精确解 $y(x)$ 是 $p+4$ 次连续可微[设 $p \geqslant 2$ 以及假设 (7-88) 为最经常使用的算子 (7-9) 和 (7-10) 所满足]. 于是我们有

$$-y(x_{n-1}) + 2y(x_n) - y(x_{n+1}) + h^2\{\beta_0 f(x_{n-1}, y(x_{n-1})) + \beta_1 f(x_n, y(x_n)) + \beta_2 f(x_{n+1}, y(x_{n+1}))\} = -C_{p+2} y^{(p+2)}(x_n) h^{p+2} + O(h^{p+4}).$$

(由于 $\beta_0+\beta_1+\beta_2=1$，常数 C_{p+2} 在现在的情形下便恒等于 C。)我们再从 (7-85) 中减去上式. 由于 $e_n = O(h^2)$，

$$f(x_n, y_n) - f(x_n, y(x_n)) = g(x_n)e_n + O(h^4).$$

此外，由 (7-88)，

$$y^{(p+2)}(x_n) = \beta_0 y^{(p+2)}(x_{n-1}) + \beta_1 y^{(p+2)}(x_n) + \beta_2 y^{(p+2)}(x_{n+1}) + O(h^2).$$

它除以 h^p，并用 $\bar{e}_n = h^{-p}e_n$ 来定义伸缩误差 $\bar{e}_n$，将所得关系式写成下面的形式：

$$-\bar{e}_{n-1} + 2\bar{e}_n - \bar{e}_{n+1} + h^2\{\beta_0\Phi_{n-1} + \beta_1\Phi_n + \beta_2\Phi_{n+1}\} = O(h^4),$$

其中

$$\Phi_n = g(x_n)\bar{e}_n - Cy^{(p+2)}(x_n),\ n = 1,2,\cdots,N-1.$$

用现在的有限差分方法求解边值问题

$$e''(x) = g(x)e(x) - Cy^{(p+2)}(x),\ e(a) = e(b) = 0, \tag{7-89}$$

并用 $\bar{e}_n$ 表示 $e(x_n)$ 的近似值，将产生同样的关系式[在右边缺少 $O(h^4)$ 这一项]。 由于问题 (7-89) 是明显地属于M类，我们可以借助定理 7.8，得到

$$\bar{e}_n = e(x_n) + O(h^2),\ n = 1,2,\cdots,N-1$$

的结论，其中 $e(x)$表示边值问题 (7-89) 的解。 因此我们证明了：

定理 7.9. 在本节开始时所叙述的假设下，误差 e_n 满足

$$e_n = h^p e(x_n) + O(h^{p+2}),\ n = 1,2,\cdots,N-1, \tag{7-90}$$

其中 $e(x)$ 表示边值问题 (7-89) 的解。

这里值得注意的是，在 (7-90) 的余项中的指数是 $p+2$ 而不是 $p+1$，这正如所期望的那样。 这个结果的实际意义是 Richardson 延迟趋于极限，如果按 §2.2-7 中所解释的那样来应用的话，将把精确度提高二个数量级(假设没有舍入误差而 h 已是充分小)。 也可见第 6 章与此有关的问题 16。

虽然在这一节的分析中假设所考虑的边值问题是属于M类的，可以相信结果 (7-90) 对并不满足这个假设的某些问题也是成立的。

7.3-3. 差分校正. 假设给定的微分方程(7-1)是线性的. 此时在(7-89)中的函数 $g(x)$ 则是已给定的微分方程 y 的系数. 在一个实际问题中，我们还不能用(7-89)来确定 $e(x)$，因为 $y(x)$是未知的，因而 $y^{(p+2)}(x)$ 也是未知的. 然而，我们在受到启示的情形下，可以如下的进行：如果(7-90)中的 $O(h^{p+2})$ 这项为零，我们便有

$$y_n = y(x_n) + h^p e(x_n). \tag{7-91}$$

因此，假设 $p = 2q$ 为偶数，而 $e(x)$ 是充分可微，利用第五章问题 32 的结果，

$$\nabla^{p+2} y_{n+q+1} = h^{p+2} y^{(p+2)}(x_n) + h^{2p+2} e^{(p+2)}(x_n) + O(h^{p+4}).$$

于是

$$y^{(p+2)}(x_n) = h^{-p-2} \nabla^{p+2} y_{n+q+1} + O(h^2).$$

因此，在关于边值问题(7-89)的一个有限差分近似中，如果用 $h^{-p-2}\nabla^{p+2} y_{n+q+1}$ 替代 $y^{(p+2)}(x_n)$ 的值，则在差分近似中只提供了一个与 $O(h^4)$ 同阶的误差，因而由定理 7.8 差分格式的解 e_n^* 与 $e(x_n)$ 的差仅为 $O(h^2)$. 按照(7-91)，值 $y_n + h^p e_n^*$ 与 $y(x_n)$ 仅相差 $O(h^{p+2})$.

$h^p e_n^*$ 所表示的量由 L. Fox [1957] 称为差分校正，在该书中对许多例子以及更为改进的方法进行了描述. 该方法也可应用于非线性方程，此时量 $g(x_n)$ 应换成 $f_y(x_n, y_n)$. 方法的严格证明依赖于下面的事实，即在适当光滑的假设下，(7-90) 能改进为

$$e_n = h^p e_p(x_n) + h^{p+2} e_{p+2}(x_n) + O(h^{p+4}), \tag{7-92}$$

其中 $e_p(x) = e(x)$，而 $e_{p+2}(x)$是 x 的某个连续微函数. 我们把(7-92)的证明以及 $e_{p+2}(x)$ 的准确的确定放到问题这一节中讨论(见问题 21).

7.4. 舍入误差的影响

7.4-1. 先验界. 由 §7.1 中描述的算法，并在一个特定的计算机上计算所得的数值 $\tilde{y}_n$，一般来说不会精确地满足关式 (7.24)，而是满足以下形式所书写出的关系式：

$$-\tilde{y}_{n+1}+2\tilde{y}_n-\tilde{y}_{n+1}+h^2\{\beta_0 f(x_{n-1},\tilde{y}_{n-1})+\beta_1 f(x_n,\tilde{y}_n)$$
$$+\beta_2 f(x_{n+1},\tilde{y}_{n+1})\}=\varepsilon_n,\quad n=1,2,\cdots,N-1. \tag{7-93}$$

我们将仍称 ε_n 为局部舍入误差。然而，必须认识到，ε_n 的出现可能不仅是由于舍入误差，也由于 Newton 方法在有限步数后就停止了，或者是方程组 (7-24) 由于其它的原因没有被精确地解出。在任何情形下，ε_n 就其实在意义可归结为舍入误差，它可以通过将值 $\tilde{y}_n$ 代入已知方程来确定。

如果假设

$$|\varepsilon_n|\leqslant\varepsilon,\quad n=1,2,\cdots,N-1, \tag{7-94}$$

用我们在证明定理 7.8 时所使用的方法，很快得到舍入误差 $r_n=\tilde{y}_n-y_n$ 的界，其结果为

$$|r_n|\leqslant\frac{(x_n-a)(b-x_n)}{2}\frac{\varepsilon}{h^2}. \tag{7-95}$$

这个界与 h^2 同阶，正如以后我们将看到的那样，它表示了当所有 ε_n 都是同号时 r_n 真正的数量级。

7.4-2. 改进的界. 我们用

$$|\varepsilon_n|\leqslant p(x_n)\varepsilon,\quad n=1,2,\cdots,N-1$$

来代替假设 (7-94)，其中 $p(x)$ 在 $a\leqslant x\leqslant b$ 上是一个非负的连续函数，并且

$$\varepsilon\leqslant K_1h^3. \tag{7-96}$$

由定理 7.8 [令 $q=1$ 和 $K=K_1\max p(x)$]，得到

$$\tilde{y}_n - y(x_n) = O(h).$$

令 $g(x) = f_y(x, y(x))$，因此得到

$$f(x_n, \tilde{y}_n) - f(x_n, y(x_n)) = g(x_n) r_n + \theta_n K_2 h^2, \quad (7-97)$$

其中 $|\theta_n| \leqslant 1$，而 K_2 是与 h 无关的常数，当 $f_{yy} = 0$ 时它为零，亦即对于线性方程来说它为零. 从 (7-85) 中减去(7-93)后，再利用 (7-97)，令 $g(x_n) = g_n$，我们得到

$$-r_{n-1} + 2r_n - r_{n+1} + h^2\{\beta_0 g_{n-1} r_{n-1} + \beta_1 g_n r_n + \beta_2 g_{n+1} r_{n+1}\}$$
$$= \varepsilon_n + \theta_n K_2 h^4,$$

或者使用矩阵的记号并用通常的办法来定义 $\mathbf{r}$，$\boldsymbol{\epsilon}$ 及 $\boldsymbol{\theta}$，我们有

$$(\mathbf{J} + h^2 \mathbf{BG})\mathbf{r} = \boldsymbol{\epsilon} + h^4 K_2 \boldsymbol{\theta}.$$

假设 (7-40) 成立，于是 $\mathbf{J} + h^2 \mathbf{BG}$ 为单调的，因而是非奇异的，我们有

$$\mathbf{r} = \mathbf{r}^{(1)} + \mathbf{r}^{(2)},$$

其中 $\mathbf{r}^{(1)}$ 称为主要舍入误差，并由

$$\mathbf{r}^{(1)} = (\mathbf{J} + h^2 \mathbf{BG})^{-1} \boldsymbol{\epsilon}$$

给出，而次要舍入误差 $\mathbf{r}^{(2)}$ 由

$$\mathbf{r}^{(2)} = h^4 K_2 (\mathbf{J} + h^2 \mathbf{BG})^{-1} \boldsymbol{\theta}$$

给出.

利用 $\mathbf{0} \leqslant (\mathbf{J} + h^2\mathbf{BG})^{-1} \leqslant \mathbf{J}^{-1}$ 和 $|\mathbf{J}^{-1}| = (b-a)^2/8h^2$，易得

$$|\mathbf{r}^{(2)}| \leqslant \frac{1}{8} h^2 K_2 (b-a)^2.$$

因此 $\mathbf{r}^{(2)} = O(h^2)$，而 $\mathbf{r}^{(1)}$ 一定被希望为 $O(h)$. 当 h 为小量时，$\mathbf{r}$ 的主要影响来自于主要舍入误差，而 $\mathbf{r}^{(2)}$ 所表示的非线性影响则可略去不计.

我们用

$$\mathbf{D} = h^{-2}(d_{mn}) = (\mathbf{J} + h^2\mathbf{BG})^{-1} \quad (7-98)$$

来定义矩阵 $\mathbf{D}$. 虽然 $\mathbf{D}$ 的元素依赖于 h, 但当 h 充分小时已知它们的元素为非负且为 $\mathbf{J}^{-1}$ 的相应元素所囿. 更精确地确定 $\mathbf{D}$ 留给下一节,我们有

$$|\mathbf{r}_m^{(1)}| \leqslant \frac{\varepsilon}{h} \sum_{n=1}^{N-1} p_n d_{nm}, \tag{7-99}$$

其中 $p_n = p(x_n)$.

7.4-3. 矩阵 D. 在这一节中, 我们将研究矩阵 $h\mathbf{D}$ 的元素 d_{nm} 当 $h \to 0$ 时的性态.

令 $g(x) = f_y(x, y(x))$, 且令函数 $s(x)$ 和 $t(x)$ 分别规定为初值问题

$$s'' = g(x)s, \quad s(a) = 0, \quad s'(a) = 1 \tag{7-100}$$

以及"终值"问题

$$t'' = g(x)t, \quad t(b) = 0, \quad t'(b) = -1 \tag{7-101}$$

的解. 我们注意到 $s(x)$ 和 $t(x)$ 的 Wronsky 行列式

$$W(x) = s'(x)t(x) - t'(x)s(x) \tag{7-102}$$

为常数,由于

$$\begin{aligned} W'(x) &= s''(x)t(x) - t''(x)s(x) \\ &= g(x)[s(x)t(x) - t(x)s(x)] = 0. \end{aligned}$$

$W(x)$ 的这个常数值异于零,因为例如

$$W(b) = -t'(b)s(b) = s(b),$$

而由 (7-100), $s(b) \geqslant b - a > 0$.

我们引入一个函数 $G(x, \xi)$, 当 $x \in [a, b]$ 和 $\xi \in [a, b]$ 时,它由

$$G(x,\xi) = \begin{cases} W^{-1}s(x)t(\xi), & x \leqslant \xi, \\ W^{-1}s(\xi)t(x), & x \geqslant \xi \end{cases} \tag{7-103}$$

确定. 函数 $G(x,\xi)$ 称为边值问题:

$$y'' = g(x)y, \quad y(a) = A, \quad y(b) = B$$

的 Green 函数, 它在这个线性边值问题的许多理论研究中起

着重要作用. 由 $G(x,\xi)$ 以及 $s(x)$ 和 $t(x)$ 的定义,容易证明下面的关系式成立:

$$G(a,\xi)=G(b,\xi)=0,\ a\leqslant\xi\leqslant b,\quad(7\text{-}104a)$$

$$\frac{d^2}{dx^2}G(x,\xi)=g(x)G(x,\xi),\ a\leqslant x\leqslant\xi,\ \xi<x\leqslant b,$$

(7-104b)

$$\lim_{x\to\xi-0}\frac{d}{dx}G(x,\xi)-\lim_{x\to\xi+0}\frac{d}{dx}G(x,\xi)=1,\ a<\xi<b.$$

(7-104c)

关系式 (7-104c) 说明 $G(x,\xi)$ 关于 x 的导数在 $x=\xi$ 处有一跳跃间断,并且在 $x=\xi$ 时左导数与右导数的差为 1. 还可以证明 $G(x,\xi)$ 为定义于 $x,\xi\in[a,b]$ 上满足 (7-104) 的三个条件的唯一函数. 作为 (7-103) 或是 (7-104) 的一个结果,还可得到

$$G(x,\xi)=G(\xi,x),\ x,\xi\in[a,b].\quad(7\text{-}105)$$

现在我们将证明:

定理 7.10. 令有限差分算子 (7-8) 的阶 p 至少为 2, 并令 $g(x)$ 在 $[a,b]$ 中二次连续可微, $g(x)\geqslant0$, 那么存在一个常数 K, 使得由 (7-98) 定义的数 d_{mn} 对所有满足 (7-32) 的 h 以及所有 $x_m,x_n\in[a,b]$ 都有

$$|d_{mn}-G(x_m,x_n)|\leqslant h^2K\quad(7\text{-}106)$$

成立.

证. 从 $(\mathbf{J}+h^2\mathbf{BG})\mathbf{D}=\mathbf{I}$, 得到

$$-d_{m-1,n}+2d_{mn}-d_{m+1,n}+h^2\{\beta_0g_{m-1}d_{m-1,n}+\beta_1g_md_{mn}$$
$$+\beta_2g_{m+1}d_{m+1,n}\}=\begin{cases}0, & m\neq n,\\ h, & m=n.\end{cases}\quad(7\text{-}107)$$

对于任何定义在 $[a,b]$ 上的函数 $z(x)$ 以及所有使 $x+h$ 及 $x-h$ 均在 $[a,b]$ 中的 x, 令

$$L[z(x);h] = -z(x-h) + 2z(x) - z(x+h)$$
$$+ h^2\{\beta_0 g(x-h)z(x-h) + \beta_1 g(x)z(x)$$
$$+ \beta_2 g(x+h)z(x+h)\}.$$

用在 §6.1-3 中使用过的方法，容易证明，如果

$$g(x)z(x) = z''(x),$$

并且 $z(x)$ 是 4 次连续可微的，便有

$$|L[z(x);h]| \leqslant h^4 A Z_4, \qquad (7\text{-}108)$$

其中 A 为一个常数，它仅依赖于 β_0, β_1, β_2, 而 Z_4 为 $|z^{\mathrm{IV}}(x)|$ 的一个界。

现在令 $0 < m < n$。假设 L 作用于 $G(x,\xi)$ 的第一个变量上，于是便有

$$L[G(x_m;x_n);h] = t(x_n)L[s(x_m);h]$$

用 S_i 及 T_i 分别表示 $|s^i(x)|$ 及 $|t^i(x)|(i = 0,1,2,\cdots)$ 的上界，因而有

$$L[G(x_m,x_n);h] = \theta_{mn}\frac{TAS_4}{W}h^4, \qquad (7\text{-}109a)$$

其中 $|\theta_{mn}| \leqslant 1$。类似地，对于 $n < m < N$，我们得到

$$L[G(x_m,x_n);h] = \theta_{mn}\frac{SAT_4}{W}h^4, \qquad (7\text{-}109b)$$

这里仍有 $|\theta_{mn}| \leqslant 1$。当 $m = n$ 时，仍然假设 L 仅作用于 G 的第一个变量，我们有

$$WL[G(x_n,x_n);h] = t(x_n)[s(x_n) - s(x_{n-1})]$$
$$+ s(x_n)[t(x_n) - t(x_{n+1})] + h^2\{\beta_0 t(x_n)g(x_{n-1})s(x_{n-1})$$
$$+ \beta_1 g(x_n)t(x_n)s(x_n) + \beta_2 s(x_n)g(x_{n+1})t(x_{n+1})\}.$$

将右端的表达式展成 h 的幂级数并使用关系式

$$gs = s'', \quad gt = t'',$$

由于 (7-104c)，我们得到

$$L[G(x_n,x_n);h]=h+\theta_n h^3\left(\frac{1}{6}+|\beta_0|\right)\frac{S_3T+ST_3}{W},$$
(7-110)

其中$|\theta_n|\leqslant 1$。

现在令

$$e_{mn}=d_{mn}-G(x_m,x_n),$$

上述结果意味着

$$-e_{m-1,n}+2e_{m,n}-e_{m+1,n}+h^2\{\beta_0 g_{m-1}e_{m-1,n}+\beta_1 g_m e_{m,n}$$
$$+\beta_2 g_{m+1}e_{m+1,n}\}=\begin{cases}\theta_{mn}K_1h^4, & m\neq n,\\ \theta_n K_2 h^3, & m=n,\end{cases}\quad (7\text{-}111)$$

其中

$$K_1=\max\left(\frac{TAS_4}{W}+\frac{SAT_4}{W}\right),\ K_2=\left(\frac{1}{6}+|\beta_0|\right)\frac{S_3T+T_3S}{W}.$$

把(7-111)看成关于$N-1$个未知量$e_{1,n},\cdots,e_{N-1,n}$具有矩阵$\mathbf{J}+h^2\mathbf{BG}$的$N$个线性方程组，由

$$0\leqslant(\mathbf{J}+h^2\mathbf{BG})^{-1}\leqslant\mathbf{J}^{-1},$$

并利用关于$\mathbf{J}^{-1}$的显式表达式(7-45)，便有

$$|e_{m,n}|\leqslant\left[\frac{(b-a)^3K_1}{8}+\frac{(b-a)K_2}{4}\right]h^2,$$

所要的结果便得到了证明.

利用(7-99)，并用积分来逼近和式，我们有

$$|r_n^{(1)}|\leqslant\frac{\varepsilon}{h^2}\{m(x_n)+O(h)\},\qquad (7\text{-}112)$$

其中

$$m(x)=\int_a^b G(x,\xi)p(\xi)d\xi.\qquad (7\text{-}113)$$

利用性质(7-108)，便可证明函数$m(x)$为边值问题

$$m''=g(x)m+p(x),\ m(a)=m(b)=0\quad (7\text{-}114)$$

的解.

7.4-4. 舍入误差的统计理论. 转到舍入误差的统计模型，现在假设局部舍入误差是 $E(r_n) = \mu_n$ 的随机变量，其中 $|\mu_n| \leqslant \mu p(x_n)$，$p(x)$ 如 §7.4-2 中所定义. 我们并不假设随机量 ϵ_n 是独立的。令

$$\text{covar}\ (\boldsymbol{\epsilon}) = E[(\boldsymbol{\epsilon} - \boldsymbol{\mu})(\boldsymbol{\epsilon}^T - \boldsymbol{\mu}^T)] = \sigma^2\mathbf{C},$$

其中矩阵 $\mathbf{C} = (C_{mn})$ 假设为已知. 由于 $\mathbf{r}^{(1)} = \mathbf{D}\boldsymbol{\epsilon}$，于是有 $E(\mathbf{r}^{(1)}) = \mathbf{D}\boldsymbol{\mu}$. 因此

$$|E(\mathbf{r}_n^{(1)})| \leqslant \frac{\mu}{h^2}\{m(x_n) + O(h)\},$$

其中 $m(x)$ 由 (7-113) 或 (7-114) 确定. 利用 §1.5-3 中所建立的法则，我们得到协方差

$$\begin{aligned}\text{covar}\ (\mathbf{r}) &= E[(\mathbf{r} - E(\mathbf{r}))(\mathbf{r}^T - E(\mathbf{r}^T))] \\ &= E[\mathbf{D}(\boldsymbol{\epsilon} - \boldsymbol{\mu})(\boldsymbol{\epsilon}^T - \boldsymbol{\mu}^T)\mathbf{D}^T] \\ &= \mathbf{D}E[(\boldsymbol{\epsilon} - \boldsymbol{\mu})(\boldsymbol{\epsilon}^T - \boldsymbol{\mu}^T)]\mathbf{D}^T \\ &= \sigma^2\mathbf{DCD}^T.\end{aligned}$$

特别感兴趣的是矩阵 covar ($\mathbf{r}$) 的对角线元素，它表示 $\mathbf{r}_n^{(1)}$ 这个量的方差. 对于这些元素，我们得到

$$\text{var}(\mathbf{r}_m^{(1)}) = \frac{\sigma^2}{h^4}\sum_{n=1}^{N-1}\sum_{k=1}^{N-1} d_{mn}c_{mk}d_{km}. \tag{7-115}$$

如果存在一个定义在 $a \leqslant \xi \leqslant b$ 和 $a \leqslant \eta \leqslant b$ 的连续函数 $c(\xi,\eta)$，使得 $c_{mn} = c(x_n, x_m)$，那么 (7-115) 中的和可以用一个重积分来近似，由此给出

$$\text{var}(r_n^{(1)}) = \frac{\sigma^2}{h^4}\{V(x_n) + O(h)\}, \tag{7-116}$$

其中

$$V(x) = \int_a^b\int_a^b G(x,\xi)c(\xi,\eta)G(\eta,x)d\xi d\eta. \tag{7-117}$$

另一方面，如果随机变量是独立的，那么矩阵 $\mathbf{C}$ 化成一个对角阵. 若它的非零元素可以用 $c_{nn} = q(x_n)$ 这样的形式

来表示，其中 $q(x)$ 是 x 的一个已知非负连续函数，则 (7-115) 中的和化成为一个简单的和，于是有

$$\operatorname{var}(r_n^{(1)}) = \frac{\sigma^2}{h^3}\{v(x_n) + O(h)\}, \qquad (7\text{-}118)$$

其中由于 (7-106)，

$$v(x) = \int_a^b G(x,\xi)^2 q(\xi) d\xi. \qquad (7\text{-}119)$$

看来将 $V(x)$ 或 $v(x)$ 用一个简单的微分方程来定义是不可能的．利用 (7-46)，容易建立不等式

$$v(x) \leqslant (b-x)^2 \int_a^x (x-\xi)^2 q(\xi) d\xi$$
$$+ (x-a)^2 \int_x^b (b-\xi)^2 q(\xi) d\xi, \qquad (7\text{-}120)$$

以及关于 $V(x)$ 的一个类似的不等式。

7.5. 问题和补充附注

§7.1

1. 求解边值问题：

$$y'' = (1+x^2)y, \quad -1 < x < 1,$$
$$y(-1) = y(1) = 1,$$

采用 4 阶方法 (7-10)，取步长 $h = 0.2$ [由于对称性，只需要考虑区间的一半]。

2. 求解边值问题：

$$y'' = 6x + y^3, \quad 0 < x < 1,$$
$$y(0) = y(1) = 0.$$

用差分算子 (7-9)，取 $h = 0.25$ [用 Newton 方法解非线性差分方程，并用略去 y^3 这一项所得到的边值问题的解作为初始近似值]。

3. 用 7.1-3 中所描述的 Gauss 算法求解方程组 $\mathbf{J}\mathbf{y}_n=\mathbf{e}_n$，其中 $\mathbf{e}_n$ 是第 n 个 $(N-1)$ 维单位向量 $(n=1, 2, \cdots, N-1)$.

4. 求解方程组

$$\mathbf{J}\mathbf{y}=\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix},$$

并证明 $\mathbf{y}=\sum_{n=1}^{N-1}\mathbf{y}_n$.

5*. 关于微分方程

$$y^{\text{IV}}+Ay''=f(x,y),\ A=\text{常数}$$

的一个明显的差分近似是由差分方程

$$\nabla^4 y_{n+2}+Ah^2\nabla^2 y_{n+1}-h^4 f_n=0$$

给出. 在差分方程

$$\begin{aligned}\nabla^4 y_{n+2}&+Ah^2\{\nabla^2 y_{n+1}+\alpha\nabla^4 y_{n+2}\}\\&=h^4\{f_n+\beta\nabla^2 f_{n+1}+\delta\nabla^4 f_{n+2}\}\end{aligned}$$

中适当定出常数 α, β, δ 来改进上述差分近似的精确度.

6. 把 7.1-3 中求解 $\mathbf{AX}=\mathbf{b}$ 的算法（其中 $\mathbf{A}$ 为三对角的）推广到 $\mathbf{A}$ 为五对角的情形，即当 $|i-j|>2$ 时 $a_{ij}=0$.

7. 考虑方程组的边值问题:

$$\mathbf{y}''=\mathbf{G}(x)\mathbf{y}+\mathbf{b}(x),\ \mathbf{y}(a)=\mathbf{l},\ \mathbf{y}(b)=\mathbf{r},$$

每一个微分方程用方法 (7-8) 来代替. 如果引进

$$(N-1)\times k$$

维复合向量 $\mathbf{y}$，其分量为 $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{N-1}$ 所产生的线性方程组可以书写成

$$\mathbf{A}\mathbf{y}=\mathbf{B}, \tag{7-121}$$

其中 $\mathbf{B}$ 为某个依赖于 $\mathbf{b}(x)$, $\mathbf{l}$ 和 $\mathbf{r}$ 的向量，而 $\mathbf{A}$ 表示复合矩

阵

$$\mathbf{A}=\begin{pmatrix} 2\mathbf{I}+h^2\mathbf{G}(x_1) & -\mathbf{I} & 0\cdots 0 \\ -\mathbf{I} & 2\mathbf{I}+h^2\mathbf{G}(x_2) & -\mathbf{I}\cdots 0 \\ & \ddots & \ddots \\ 0 & & \ddots \\ & & -\mathbf{I} \quad 2\mathbf{I}+h^2\mathbf{G}(x_{N-1}) \end{pmatrix} \tag{7-122}$$

($\mathbf{I}=k\times k$(单位矩阵). 建立一个关于方程组(7-121)数值解的算法,它完全是以 $k\times k$ 块矩阵进行计算的.

8. 对边值条件:

$$y'(a)-cy(a)=A,\ y'(b)+dy(b)=B, \tag{7-123}$$

证明类似于定理 7.1 的结论,其中 $c\geqslant 0$, $d\geqslant 0$, $c+d>0$ [用 $y(a)=\alpha$ 作为一个参数,并证明

$$y_\alpha(x,\alpha)\geqslant 1+c(x-a), y'_\alpha(x,\alpha)\geqslant c].$$

§7.2.

9. (a) 证明单位矩阵是可约的.

(b) 令 $\mathbf{A}=(a_{ij})$ 和 $\mathbf{B}=(b_{ij})$ 是两个 n 阶矩阵. 如果 $\mathbf{A}$ 是不可约的,且 $a_{ij}\neq 0$ 意指 $b_{ij}\neq 0$, 则 $\mathbf{B}$ 是可约的.

10. 一个三对角矩阵 $\mathbf{A}=(a_{ij})$ 是可约的,当且仅当对某些 i 有 $a_{i,i-1}a_{i-1,i}=0$. 同时用定义及定理 7.2 来证明这一结论.

11. 令 α, β 及 $\gamma\neq 0$, 2×2 矩阵

$$\text{(a)}\begin{pmatrix}\alpha & \beta\\ \gamma & 0\end{pmatrix},\ \text{(b)}\begin{pmatrix}0 & \beta\\ \gamma & 0\end{pmatrix},\ \text{(c)}\begin{pmatrix}0 & 0\\ \gamma & 0\end{pmatrix}$$

中哪一个是可约的?

12. 假设所有矩阵 $\mathbf{G}(x_n)(n=1,2,\cdots,N-1)$ 满足定理 7.4 的条件,则由(7-122)所定义的矩阵 $\mathbf{A}$ 是非奇异的.

13. 由微分方程 $y'' = g(x)y + h(x)$ 在边界条件 (7-123) 下，对于充分小的 h 进行离散化所得到的线性方程组的矩阵满足定理 7.4 的假设，因而是单调的。

14. 用 Newton 法求解方程

$$x^2 - a = 0, \tag{7-124}$$

其中 $a > 0$，这里 $x_0 > 0$。

(a) 证明 Kantorovich 的定理 7.6 的假设是满足的；

(b) 证明解 $x = \sqrt{a}$ 位于由 (7-65) 定义的区域的边界上；

(c) 证明逐次近似值满足关系式：

$$\frac{x_n - \sqrt{a}}{x_n + \sqrt{a}} = \left(\frac{x_0 - \sqrt{a}}{x_0 + \sqrt{a}}\right)^{2^n},$$

并比较真正的误差 $x_n - \sqrt{a}$ 与由 (7-66) 所给出的估计。

15. 关于定理 5.4 中描述的解 $x = f(x)$ 的迭代过程，其收敛的阶如何?

16*. 改进的 Newton 过程的收敛性。 令 $\mathbf{z}^{(0)}, \mathbf{z}^{(1)}, \cdots$ 表示用

$$\mathbf{z}^{(\nu+1)} = \mathbf{z}^{(\nu)} - [\mathbf{A}(\mathbf{z}^{(0)})]^{-1}\boldsymbol{\varphi}(\mathbf{z}^{(\nu)}), \quad \nu = 0, 1, 2, \cdots \tag{7-125}$$

改进的 Newton 方法求解 $\boldsymbol{\varphi}(\mathbf{y}) = 0$ 所得到的逐次近似值 [泛函矩阵 $\mathbf{A}(\mathbf{z})$ 只求一次逆]，如果满足定理 7.6 的假设而且

$$h_0 < \frac{1}{2},$$

证明

$$|\mathbf{z}^{(\nu)} - \mathbf{y}^*| \leqslant q^{\nu-1}|\mathbf{z}^{(1)} - \mathbf{y}^*|, \tag{7-126}$$

其中 $q = 1 - \sqrt{1 - 2h_0} < 1$，并由此得出结论

$$\lim_{\nu\to\infty} \mathbf{z}^{(\nu)} = \mathbf{y}^*$$

(见 Kantorovich [1948], p. 181).

17. 把改进的 Newton 方法用于 (7-124)，证明估计式 (7-126)。你能否构造一个函数 $f(x)$，对它来说估计式是明显的?

§7.3.

18. 对边值问题 $y'' = y - 2$，$y(0) = 2$，$y(1) = 0$ 解析地进行差分计算，使用 (7-9)，并证明定理 7.9 的结论.

19. 对边值问题(不属于 M 类)

$$y'' = y, \quad y(0) = 0, \quad y'(1) = 1,$$

使用算子 (7-9) 解析地进行差分计算，并决定误差的阶，如果 $y'(1)$ 是用 (a) $h^{-1}(y_N - y_{N-1})$ 来近似的; (b) 用

$$(2h)^{-1}(y_{N+1} - y_{N-1})$$

来近似的.

20. 用差分校正改进问题 2 中得到的数值结果，以微分来获得所要求的 y^{VI} 值，如果必要的话外推前面获得的 $y'' = (1 + x^2)y$ 值. 并对以此得到的值与精确解的 6 阶导数进行比较.

21*. 差分校正的证明. 令 $L[y(x);h]$ 表示与差分表达式 (7-8) 相关的差分算子，如果 $y(x)$ 是充分可微的并且

$$L[y(x);h] = C_{p+2}y^{(p+2)}(x)h^{p+2} + C_{p+4}y^{(p+4)}(x)h^{p+4} + O(h^{p+6}).$$

证明取 $e_{p+2}(x) = \eta(x)$ 时 (7-92) 成立，当

(a) $p = 2$，$\beta_0 = \beta_2 = 0$，而 $\eta(x)$ 是边值问题

$$\eta'' = g(x)\eta + \frac{1}{2} f_{yy}(x, y(x))[e(x)]^2 + C_6 y^{VI}(x) + C_4 e^{IV}(x),$$

$$\eta(a) = \eta(b) = 0$$

的解;

(b) $p=4$，$\beta_0=\beta_2$，而 $\eta(x)$ 是

$$\eta''=g(x)\eta-(C_8+\beta_0C_6)y^{\mathrm{VIII}}(x),\quad \eta(a)=\eta(b)=0$$

的解。

22. 蕴含定理 (Collatz [1960], p. 200). 若对于一个属于 M 类的边值问题，满足边界条件二个函数 $y_1(x)$ 和 $y_2(x)$ 发现使得

$$-y_1''(x)+f(x,y_1(x))\leqslant 0\leqslant -y_2''(x)+f(x,y_2(x))$$

$$a\leqslant x\leqslant b,$$

那么精确解 $y(x)$ 满足

$$y_1(x)\leqslant y(x)\leqslant y_2(x),\quad a\leqslant x\leqslant b.$$

§7.4.

23. 证明当 $g(x)=1$ 时，

$$G(x,\xi)=\frac{\sinh[\min(x,\xi)-a]\sinh[b-\max(x,\xi)]}{\sinh(b-a)}$$

以及假设 $p(x)=q(x)=1$ 时，

$$m(x)=1-\frac{\cosh\left[\frac{1}{2}(b+a)-x\right]}{\cosh\left[\frac{1}{2}(b-a)\right]},$$

$$v(x)=\frac{\sinh(x-a)\sinh(b-x)}{\sinh(b-a)}-\frac{(x-a)\sinh^2(b-x)+(b-x)\sinh^2(x-a)}{2\sinh^2(b-a)}.$$

24. 证明关系式 (7-104)。

25. 证明边值问题

$$y''=g(x)y+p(x),\quad y(a)=A,\quad y(b)=B$$

的解可以表成

$$y(x)=\frac{A}{t(a)}t(x)+\frac{B}{s(b)}s(x)+\int_a^b G(x,\xi)p(\xi)d\xi.$$

26. 用 $G_i(x,\xi)(i=1,2)$ 表示与边值问题 $y''=g_i(x)y$, $y(a)=A$, $y(b)=B$（其中 $0\leqslant g_1(x)\leqslant g_2(x)$, $x\in[a,b]$）有关的 Green 函数，证明 $0\leqslant G_2(x,\xi)\leqslant G_1(x,\xi)$, x, $\xi\in[a,b]$.

注

§7.1-2. 其它求解边值问题的离散变量方法为 Gavurin [1949], Stüssi [1950], Mikelazde [1953a], Karpilovsaya [1953], Glinskaya 和 Mysovskih [1954], Adachi [1955], Goodman 和 Lance [1956], Tihonov 和 Samarskii [1956], Ridley [1957], Fox [1957] 所提出。 Sturm-Liouville 问题为 Farrington, Gregory 和 Taub [1957] 以及 Wachspress [1960] 所讨论. 具有非均匀网格步长的差分方程为 Brown [1960] 所研究。

§7.1-3. 这儿所用的 Gauss 算法的特殊形式（有时略加改变）是 Fox [1957], Anonymous [1957], Douglas [1959], Forsythe 和 Wasow [1960], p. 103, Wachspress [1960] 所讨论. 这个方法容易为矩阵 **A** 具有多于三个非零对角线的矩阵 **A** 所采用，见 Rutishauser [1958], Conte 和 Dames [1958] 以及问题 6.

§7.1-4. 用 Newton 方法求解由非线性边值问题中引出的差分方程的数值试验为 Ehrlich [1960] 和 Mercer [1960] 所报导。关于另一个迭代方法，见 Pope [1960].

§7.2-1. 定理 7.2 为 Varga [1959] 所说明.

§7.2-2. 在单调矩阵的定义和某些讨论中，我们按照 Collatz [1960], p. 43 来处理.

§7.2-4. 定理 7.6 的证明是对有限维的情形，是用一个特殊范数的 Kantorovich 的原始证明（Kantorovich [1948]）的一种小改进.

§7.3-1. 关于更一般的边值问题的有限差分近似误差界，

可以在 §7.1-2 中列出的文献以及在 Schröder [1956, 1958] 中找到。Mitchell [1959] 曾提出关于不属于 M 类的边值问题，有限差分方法的固有困难。

§7.4. 在解具有三对角矩阵的线方程组过程中，舍入误差的传播为 Douglas [1959]，Wachspress [1960]，Lowan [1960] 所讨论。

参 考 文 献

我们试图把最近十年来在所有的书或者已出版的期刊中所出现的有关文章编成尽可能完善的目录，包括作者所注意到并认为有关的教学笔记、手稿以及未出版的研究报告，也包括 1950 年前的作者认为有持久重要性的文章。在 1952 年前的更完全的文献目录可在 Milne [1953] 的书中找到，1932 年前的文献可见 Beunett, Milne 和 Batemen [1956]。

不直接涉及到常微分方程离散变量方法的辅助参考资料均用星号标出。

Adachi, R. [1955]: A method on the numerical solution of some differential equations. *Kumamoto J. Sci. Ser. A*, 2, 244–252.

*Ahlfors, L. V. [1953]: *Complex Analysis*, McGraw-Hill, New York.

Albrecht, J. [1955]: Beiträge zum Runge-Kutta-Verfahren. *Z. angew. Math. Mech.*, 35, 100–110.

Alonso, R. [1960]: A starting method for the three-point Adams predictor-corrector method. *J. Assoc. Comp. Mach.*, 7, 176–180.

Anderson, W. H. [1960]: The solution of simultaneous ordinary differential equations using a general purpose digital computer. *Comm. Assoc. Comp. Mach.*, 3, 355–360.

Anonymous [1957]: Modern Computing Methods. *Notes Appl. Sci. No.* 16, National Physical Laboratory, London.

Antosiewicz, H. A., and Walter Gautschi [1951]: Numerical methods in ordinary differential equations, to appear in *Survey of Numerical Analysis*, J. Todd (ed.), McGraw-Hill, New York.

Artemov, G. A. [1955]: On a modification of Caplygin's method for systems of ordinary differential equations of the first order. *Dokl. Akad. Nauk SSSR* (N. S.), 101, 197–200.

Azbelev, N. V. [1952]: On an approximate solution of ordinary differential equations based upon S. A. Caplygin's method. *Dokl. Akad. Nauk SSSR* (N. S.), 83, 517–519

——— [1953]: On the limits of applicability of S. A. Caplygin's theorem. *Dokl. Akad. Nauk SSSR* (N. S.), 89, 589–591.

——— [1955]: On the extension of Caplygin's method beyond the limit of applicability of the theorem on differential inequalities. *Dokl. Akad. Nauk SSSR* (N. S.), 102, 429–430.

Babkin, B. N. [1948]: Angenäherte Lösung gewöhnlicher Differentialgleichungen beliebiger Ordnung mit der Methode der schrittweisen Annäherungen auf Grund eines Satzes von S. A. Caplygin über Differentialgleichungen. *Dokl. Akad. Nauk SSSR* (2), 59, 419–422.
——— [1949]: On a modification of a method of S. A. Caplygin for approximate integration. *Dokl. Akad. Nauk. SSSR* (N. S.), **67**, 213–216.
——— [1954]: Approximate integration of systems of ordinary differential equations of the first order by the method of S. A. Caplygin. *Izv. Akad. Nauk SSSR Ser. Mat.*, **18**, 477–484.
Bahvalov, N. S. [1955a]: On the estimation of the error in the numerical integration of differential equations by the Adams extrapolation method. *Dokl. Akad. Nauk SSSR* (N. S.), **104**, 683–686.
——— [1955b]: Some remarks concerning the numerical integration of differential equations by the method of finite differences. *Dokl. Akad. Nauk SSSR* (N. S.), **104**, 805–808.
Balakin, V. B [1959]: Bilateral approximations to the solution of the equation $y^{(n)} = f(x, y)$. *Ukrain. Mat. Z.*, **11**, 203–207.
Bashforth, F., and J. C. Adams [1883]: *Theories of Capillary Action*, Cambridge Univ. Press.
Bennett, A. A., W. E. Milne, and H. Bateman [1956]: *Numerical Integration of Differential Equations*, Dover, New York.
Bieberbach, L. [1944]: *Theorie der Differentialgleichungen*, Dover, New York.
——— [1951]: On the remainder of the Runge-Kutta formula in the theory of ordinary differential equations. *Z. angew. Math. Physik*, **2**, 233–248.
*Birkhoff, G., and S. MacLane [1953]: *A Survey of Modern Algebra*, rev. ed., Macmillan, New York.
Blanch, G. [1952]. On the numerical solution of equations involving differential operators with constant coefficients. *Math. Tables Aids Comput.*, **6**, 219–223.
——— [1957]: Criteria for the choice of an integration formula. Talk delivered at USE meeting, Dayton.
Blum, E. K. [1957]: A modification of the Runge-Kutta fourth-order method. *Numerical Note N-N 80*, Ramo-Wooldridge Corp., Los Angeles.
Brock, P., and F. J. Murray [1952]: The use of exponential sums in step by step integration. *Math. Tables Aids Comput.*, **6**, 63–78, 138–150.
Brodskii, M. L. [1953]: Asymptotic estimates of the errors in the numerical integration of systems of ordinary differential equations by difference methods. *Dokl. Akad. Nauk SSSR* (N. S.), **93**, 599–602.
Brouwer, D. [1937]: On the accumulation of errors in numerical integration. *Astronomical J.*, **46**, 149–153.
Brown, R. R. [1960]: Solution of boundary value problems using nonuniform grids. Ph.D. dissertation, Univ. California, Los Angeles.
Bückner, H. [1952]: Ueber eine Näherungslösung der gewöhnlichen linearen Differentialgleichung 1. Ordnung. *Z. angew. Math. Mech.*, **22**, 143–152.
Budak, B. M. [1956]: On the method of straight lines for certain boundary problems. *Vestnik Moskov. Univ. Ser. Mat. Meh. Astr. Fiz. Him.*, **11**, 3–12.
——— and A. D. Gorbunov [1959]: Stability of calculation processes in the solution of the Cauchy problem for the equation $dy/dx = f(x, y)$ by multipoint difference methods. *Dokl. Akad. Nauk SSSR*, **124**, 1191–1194.
Bukovics, E. [1950]: Eine Verbesserung und Verallgemeinerung des Verfahrens von Blaess zur numerischen Integration gewöhnlicher Differentialgleichungen. *Oesterreich. Ing.-Archiv*, **4**, 338–349.

——— [1953]: Beiträge zur numerischen Integration, I, II. *Monatshefte der Math.*, 57, 217–245; 57, 333–350.
——— [1954]: Beiträge zur numerischen Integration, III. *Monatshefte der Math.*, 58, 258–265.
Capra, V. [1956]: Nuove formule per l'integrazione delle equazioni differenziali ordinarie del 1° e del 2° ordine. *Univ. e Politec. Torino. Rend. Sem. Mat.*, 16, 301–359.
——— [1957]: Valutazione degli errori nella integrazione numerica dei systemi di equazioni differenziali ordinarie. *Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Nat.*, 91, 188–203.
Carr, J. W. [1957]: Lecture notes on ordinary differential equations. Applications of Advanced Numerical Analysis to Digital Computer Problems, *Summer session*, Univ. Michigan.
——— [1958]: Error bounds for the Runge-Kutta single-step integration process. *J. Assoc. Comp. Mach.*, 5, 39–44.
Cauchy, A. [1840]: Mémoire sur l'intégration des équations différentielles. *Oeuvres complètes*, II° série, tome 11, 399–465.
Cernysenko, E. A. [1958]: On a method of approximate solution of Cauchy's problem for ordinary differential equations. *Ukrain. Mat. Z.*, 10, 89–100.
Certaine, J. [1960]: The solution of ordinary differential equations with large time constants, in *Mathematical Methods for Digital Computers*, A. Ralston and H. Wilf (eds.), Wiley, New York, pp. 128–132.
Ceschino, F. [1954]: Critère d'utilisation du procédé de Runge-Kutta. *C. R. Acad. Sci. Paris*, 238, 986–988, 1553–1555.
——— [1956]: L'intégration approchée des équations différentielles. *C. R. Acad. Sci. Paris*, 243, 1478–1479.
——— and J. Kuntzmann [1958]: Impossibilité d'un certain type de formule d'intégration approchée à pas liés. *Chiffres*, 1, 95–101.
——— [1960]: Faut-il passer à la forme canonique dans les problèmes différentielles de conditions initiales? *Information Processing*, UNESCO, Paris, pp. 33–36.
*Coddington, E. A., and N. Levinson [1955]: *Theory of Ordinary Differential Equations*, McGraw-Hill, New York.
Collatz, L. [1949]: *Eigenwertaufgaben mit technischen Anwendungen*, Akad. Verlagsgesellschaft, Leipzig.
——— [1953]: Ueber die Instabilität beim Verfahren der zentralen Differenzen für Differentialgleichungen zweiter Ordnung. *Z. angew. Math. Physik*, 4, 153–154.
——— [1960]: *The Numerical Treatment of Differential Equations*, 3rd ed., Springer, Berlin.
——— and R. Zurmühl [1942]: Zur Genauigkeit verschiedener Integrationsverfahren bei gewöhnlichen Differentialgleichungen. *Ing.-Arch.*, 13, 34–36.
Conte, S. D. [1958]: Stable operators in the numerical solution of second order differential equations. *Numerical Analysis Note N-N 112*, Space Technology Laboratories, Los Angeles.
——— and R. T. Dames [1958]: An alternating direction method for solving the biharmonic equation. *Math. Tables Aids Comput.*, 12, 198–204.
Conte, S. D., and R. F. Reeves [1956]: A Kutta third-order procedure for solving differential equations requiring minimum storage. *J. Assoc. Comp. Mach.*, 3, 22–25.
Cowell, P. H., and A. C. D. Crommelin [1910]: Investigation of the motion of Halley's comet from 1759 to 1910. Appendix to Greenwich Observations for 1909, Edinburgh, p. 84.
*Cramér, H. [1946]: *Mathematical Methods of Statistics*, Princeton Univ. Press.
Dahlquist, G. [1951]: Fehlerabschätzungen bei Differenzenmethoden zur numerischen

Integration gewöhnlicher Differentialgleichungen. *Z. angew. Math. Mech.*, **31**, 239–240.
——— [1956]: Convergence and stability in the numerical integration of ordinary differential equations. *Math. Scand.*, **4**, 33–53.
——— [1959]: Stability and error bounds in the numerical integration of ordinary differential equations. *Trans. Roy. Inst. Technol., Stockholm*, Nr. 130.
Dettmar, H. K., and A. Schlüter [1958]: Praktische Lösung von Eigenwertaufgaben vom Hartree-Fockschen Typus. *Z. Angew. Math. Mech.*, **38**, 220–236.
de Vogelaere, R. [1955]: A method for the numerical integration of differential equations of second order without explicit first derivatives. *J. Res. Nat. Bur. Standards*, **54**, 119–125.
——— [1957]: On a paper of Gaunt concerned with the start of numerical solutions of differential equations. *Z. angew. Math. Physik*, **8**, 151–156.
Douglas, Jim [1956]: On the error in analogue solutions of heat conduction problems. *Quart. J. Appl. Math.*, **14**, 333–335.
——— [1959]: Round-off error in the numerical solution of the heat equation. *J. Assoc. Comput. Mach.*, **6**, 48–58.
Ehrlich, L. [1960]: Experience with numerical methods for a boundary value problem. *Technical Note N-N 141*, Space Technology Laboratories, Los Angeles.
Eltermann, H. [1955]: Fehlerabschätzungen bei näherungsweiser Lösung von Systemen von Differentialgleichungen erster Ordnung. *Math. Z.*, **62**, 469–501.
Euler, L. [1913]: *Opera omnia, series prima*, Vol. 11, Leipzig and Berlin.
——— [1914]: *Opera omnia, series prima*, Vol. 12, Leipzig and Berlin.
Farrington, C. C., R. T. Gregory, and A. H. Taub [1957]: On the numerical solution of Sturm-Liouville differential equations. *Math. Tables Aids Comput.*, **11**, 131–150.
Fehlberg, E. [1958]: Eine Methode zur Fehlerverkleinerung beim Runge-Kutta-Verfahren. *Z. angew. Math. Mech.*, **38**, 421–426.
——— [1960]: Neue genauere Runge-Kutta-Formeln für Differentialgleichungen zweiter Ordnung. *Z. angew. Math. Mech.*, **40**, 252–259.
——— [1961]: Numerisch stabile Interpolationsformeln mit günstiger Fehlerfortpflanzung für Differentialgleichungen erster und zweiter Ordnung. *Z. angew. Math. Mech.*, **41**, 101–110.
*Feller, W. [1957]: *An Introduction to Probability Theory and Its Applications*, 2nd ed., Wiley, New York.
Forsythe, G. E. [1959]: Note on rounding-off errors. *SIAM Rev.*, **1**, 66–67.
——— and W. Wasow [1960]: *Finite-Difference Methods for Partial Differential Equations*, Wiley, New York.
Fort, T. [1948]: *Finite Differences and Difference Equations in the Real Domain*, Oxford, Univ. Press.
Fox, L. [1957]: *The Numerical Solution of Two-Point Boundary Problems in Ordinary Differential Equations*, Oxford Univ. Press.
——— [1960]: Some numerical experiments with eigenvalue problems in ordinary differential equations, in *Boundary Problems in Differential Equations*, R. E. Langer (ed.), Univ. of Wisconsin Press, Madison, pp. 243–255.
——— and A. R. Mitchell [1957]: Boundary-value techniques for the numerical solution of initial value problems in ordinary differential equations. *Quart. J. Mech. Appl. Math.*, **10**, 232–243.
Franklin, J. [1952]: On the method of successive approximations. *Tech. Rept. No. 7*, Appl. Math. and Statist. Lab., Stanford Univ.
——— [1959]: Numerical stability in digital and analog computation for diffusion problems. *J. Math. Phys.*, **37**, 305–315

Frei, T. [1954]: Anwendung der Momente der Integralkurven zur numerischen Lösung von Differentialgleichungen. *Magyar Tud. Akad. Alkalm. Mat. Int. Közl.*, **2**, 395–414.

Fricke, A. [1949]: Ueber die Fehlerabschätzung des Adamsschen Verfahrens zur Integration gewöhnlicher Differentialgleichungen erster Ordnung. *Z. angew. Math. Mech.*, **29**, 165–178.

Gaier, D. [1956]: Ueber die Konvergenz des Adamsschen Extrapolationsverfahrens. *Z. angew. Math. Mech.*, **36**, 230.

Galler, B. A., and D. P. Rozenberg [1960]: A generalization of a theorem of Carr on error bounds for Runge-Kutta procedures. *J. Assoc. Comp. Mach.*, **7**, 57–60.

Garfinkel, B. [1954]: On the choice of mesh in the integration of ordinary differential equations. *Rept. No. 907*, Ballistics Research Labs., Aberdeen Proving Ground, Md.

Garwick, J. V. [1955]: The solution of boundary-value problems by step-by-step methods. *Arch. Math. Naturw.*, **52**, 1–67.

Gaunt, J. A. [1927]: The deferred approach to the limit, II—Interpenetrating lattices. *Trans. Roy. Soc. Lond.*, **226**, 350–361.

Gautschi, Walter [1955]: Ueber den Fehler des Runge-Kutta-Verfahrens für die numerische Integration gewöhnlicher Differentialgleichungen *n*-ter Ordnung. *Z. angew. Math. Physik.*, **6**, 456–461.

Gavurin, M. K. [1949]: On a method of numerical integration of homogeneous linear differential equations convenient for mechanisation of the computation. *Trudy Mat. Inst. Steklov*, **28**, 152–156.

Gill, S. [1951]: A process for the step-by-step integration of differential equations in an automatic digital computing machine. *Proc. Cambridge Philos. Soc.*, **47**, 96–108.

Glinskaya, N. N., and I. P. Mysovskih [1954]: On the numerical solution of a boundary problem for a non-linear ordinary differential equation. *Vestnik Leningrad Univ.*, **9**, 49–54.

Goldberg, S. [1958]: *Introduction to Difference Equations*, Wiley, New York.

Goodman, T. R., and G. N. Lance [1956]: The numerical integration of two-point boundary value problems. *Math. Tables Aids Comput.*, **10**, 82–86.

Gorn, S., and R. Moore [1953]: Automatic error control—the initial value problem in ordinary differential equations. *Rept. No. 893*, Ballistic Research Labs., Aberdeen Proving Ground, Md.

Gray, H. J., Jr. [1955]: Propagation of truncation errors in the numerical solution of ordinary differential equations by repeated closures. *J. Assoc. Comp. Mach.*, **2**, 5–17.

Hamel, G. [1949]: Zur Fehlerabschätzung bei gewöhnlichen Differentialgleichungen erster Ordnung. *Z. angew. Math. Mech.*, **29**, 337–341.

Hammer, P. C., and T. W. Hollingsworth [1955]: Trapezoidal methods of approximating solutions of differential equations. *Math. Tables Aids Comput.*, **9**, 92–96.

Hamming, R. W. [1959]: Stable predictor-corrector methods for ordinary differential equations. *J. Assoc. Comp. Mach.*, **6**, 37–47.

Hanson, J. W. [1960]: A numerical example of integration in a plane without perturbations using Cowell's method. *Math. Note No.* 3, Computation Center, Univ. of North Carolina.

Hartree, D. R. [1952]: *Numerical Analysis*, Oxford Univ. Press.

Henrici, P. [1960]: Theoretical and experimental studies on the accumulation of error in the numerical solution of initial value problems for systems of ordinary differential equations. *Information Processing*, UNESCO, Paris, pp. 36–44.

——— [1961]: The propagation of round-off error in the numerical solution of initial value problems involving ordinary differential equations of the second order.

Proceedings of a Symposium on the Numerical Treatment of Ordinary Differential Equations, Integral Equations, and Integro-differential Equations, Rome.

Herrick, S. [1951]: Step-by-step integration of $\ddot{x} = f(x, y, z, t)$ without a "corrector." *Math. Tables Aids Comput.*, **5**, 61–67.

Heun, K. [1900]: Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen. *Z. Math. Physik*, **45**, 23–38.

Hildebrand, F. B. [1956]: *Introduction to Numerical Analysis*, McGraw-Hill, New York.

*Hoel, P. G. [1954]. *Introduction to Mathematical Statistics*, 2nd ed., Wiley, New York.

Hull, T. E., and W. A. J. Luxemburg [1960]: Numerical methods and existence theorems for ordinary differential equations. *Num. Math.*, **2**, 30–41.

Hull, T. E., and A. C. R. Newbery [1959]: Error bounds for a family of three-point integration procedures. *J. Soc. Indust. Appl. Math.*, **7**, 402–412.

——— [1961]: Integration procedures which minimize propagated errors. *J. Soc. Indust. Appl. Math.*, **9**, 31–47.

Huskey, H. D. [1949]: On the precision of a certain procedure of numerical integration. With an appendix by Douglas R. Hartree. *J. Research Nat. Bur. Standards*, **42**, 57–62.

Huta, A. [1956]: Une amélioration de la méthode de Runge-Kutta-Nyström pour la résolution numérique des équations différentielles du premier ordre. *Acta Fac. Nat. Univ. Comenian. Math.*, **1**, 201–224.

——— [1957]: Contribution à la formule de sixième ordre dans la méthode de Runge-Kutta-Nyström. *Acta Fac. Nat. Univ. Comenian. Math.*, **2**, 21–24.

Ionescu, D. V. [1954]: Une généralisation d'une propriété intervenant dans la méthode de Runge-Kutta pour l'intégration numérique des équations différentielles. *Acad. Rep. Populare Române, bul. Stint. Ser. Mat. Fiz.*, **6**, 229–241.

——— [1956]: Une généralisation d'une propriété intervenant dans la méthode de Runge-Kutta pour l'intégration numérique des équations différentielles. *Acad. Rep. Populare Române, Bul. Stint. Ser. Mat. Fiz.*, **8**, 67–100.

Jackson, J. [1924]: Note on the numerical integration of $d^2x/dt^2 = f(x, t)$. *Roy. Ast. Soc. Monthly Notices*, **84**, 602–606.

Jacobsen, L. S. [1952]: On a general method of solving second-order ordinary differential equations by phase-plane displacements. *J. Appl. Mech.*, **19**, 543–553.

*Kaluza, T. [1928]: Ueber die Koeffizienten reziproker Potenzreihen. *Math. Z.*, **28**, 161–170.

Kamke, E. [1943]: *Differentialgleichungen, Lösungsmethoden und Lösungen*, Akademische Verlagsgesellschaft, Leipzig.

——— [1947]: *Differentialgleichungen reeller Funktionen*, Dover, New York.

Kantorovich, L. V. [1948]: Functional analysis and applied mathematics. *Uspekhi Mat. Nauk*, **3**, 89–185. Translated by C. D. Benster and edited by G. E. Forsythe as Nat. Bur. Standards *Rept. No. 1509*.

Karpilovskaya, E. B. [1953]: On the convergence of an interpolation method for ordinary differential equations. *Uspekhi Mat. Nauk* (N. S.), **8**, 111–118.

Keitel, G. H. [1956]: An extension of Milne's three-point method. *J. Assoc. Comp. Mach.*, **3**, 212–222.

Kopal, Z. [1958]: Operational methods in numerical analysis based on rational approximations. *On Numerical Approximation, Proceedings of a Symposium*, R. E. Langer (ed.), Univ. of Wisconsin Press, Madison, pp. 25–43.

Korganoff, A. [1958]: Sur les formules d'intégration numérique des équations différentielles donnant une approximation d'ordre élevé. *Chiffres*, **1**, 171–180.

Kuntzmann, J. [1953]: Remarques sur la méthode de Runge-Kutta. *C. R. Acad. Sci. Paris*, **242**, 2221–2223.

——— [1959a]: Deux formules optimales du type de Runge-Kutta. *Chiffres*, 2, 21–26.
——— [1959b]: Evaluation de l'erreur sur un pas dans les méthodes à pas séparés. *Chiffres*, 2, 97–102.
Kutta, W. [1901]: Beitrag zur näherungsweisen Integration totaler Differentialgleichungen. *Z. Math. Phys.*, 46, 435–453.
Lax, P. D., and R. D. Richtmyer [1956]: Survey of the stability of linear finite difference equations. *Comm. Pure Appl. Math.*, 9, 267–293.
Levy, H., and E. A. Baggott [1934]: *Numerical Studies in Differential Equations*, Vol. 1, Watts, London. Reprinted under the title *Numerical Solutions of Differential Equations*, Dover, New York, 1950.
Liniger, W. [1957]: Zur Stabilität der numerischen Integrationsmethoden für Differentialgleichungen. Thesis, Univ. Lausanne.
*Loève, M. [1955]: *Probability Theory. Foundation. Random Sequences.* Van Nostrand, New York.
Lotkin, M. [1951]: On the accuracy of Runge-Kutta's method. *Math. Tables Aids Comput.*, 5, 128–133.
——— [1952]: A new integration procedure of high accuracy. *J. Math. Physics*, 31, 29–34.
——— [1954]: The propagation of error in numerical integrations. *Proc. Amer. Math. Soc.*, 5, 869–887.
——— [1955]: On the improvement of accuracy in integration. *Quart. Appl. Math.*, 13, 47–54.
——— [1956]: A note on the midpoint rule of integration. *J. Assoc. Comp. Mach.*, 3, 208–211.
Loud, W. S. [1949]: On the long-run error in the numerical solution of certain differential equations. *J. Math. Phys.*, 28, 45–49.
Lowan, A. N. [1960]: On the propagation of errors in the inversion of certain tridiagonal matrices. *Math. Comput.*, 14, 333–338.
Löwdin, P. O. [1952]: On the numerical integration of ordinary differential equations of the first order. *Quart. Appl. Math.*, 10, 97–111.
Lozinskii, S. M. [1953]: Estimate of the error of an approximate solution of a system of ordinary differential equations. *Dokl. Akad. Nauk SSSR* (N. S.), 92, 225–228.
Luzin, N. N. [1951]: On the method of approximate integration of academician S. A. Caplygin. *Uspehi Matem. Nauk* (N. S.), 6, 3–27.
Martin, D. W. [1958]: Runge-Kutta methods for integrating differential equations on high speed digital computers. *The Computer J.*, 1, 118–123.
Matthieu, P. [1951]: Ueber die Fehlerabschätzung beim Extrapolationsverfahren von Adams. I. Gleichungen 1. Ordnung. *Z. angew. Math. Mech.*, 31, 356–370.
——— [1953]: Ueber die Fehlerabschätzung beim Extrapolationsverfahren von Adams. II. Gleichungen zweiter und höherer Ordnung. *Z. angew. Math. Mech.*, 33, 26–41.
Mercer, R. J. [1960]: Boundary value solutions of trajectory problems. *Tech. Note PA-2399-01/2*, Space Technology Laboratories, Los Angeles.
Mikeladze, M. S. [1953a]: Numerical solution of a system of differential equations. Application of the method to the computation of rotating shells. *Akad. Nauk SSSR, Prikl. Mat. Meh.*, 17, 382–386.
——— [1953b]: Numerical solution of boundary problems for non-linear ordinary differential equations. *Soobsc. Akak. Nauk Gruzin. SSR*, 14, 133–137.
Mikulaschkova, R. [1957]: Rounding-off error in numerical calculation from the point of view of statistics. *Pokroky Mat. Fys. Astr.*, 2, 697–707.
*Miller, J. C. P. [1955]: *Tables of Weber Parabolic Cylinder Functions, Introduction.* Her Majesty's Stationery Office, London.

Milne, W. E. [1926]: Numerical integration of ordinary differential equations. *Amer. Math. Monthly*, **33**, 455–460.

——— [1949]: A note on the numerical integration of differential equations. *J. Research Nat. Bur. Standards*, **43**, 537–542.

——— [1950]: Note on the Runge-Kutta method. *J. Research Nat. Bur. Standards*, **44**, 549–550.

——— [1953]: *Numerical Solution of Differential Equations*, Wiley, New York.

——— and R. R. Reynolds [1959]: Stability of a numerical solution of differential equations. *J. Assoc. Comp. Mach.*, **6**, 196–203.

——— [1960]: Stability of a numerical solution of differential equations. II. *J. Assoc. Comp. Mach.*, **7**, 46–56.

Milne-Thomson, L. M. [1933]: *The Calculus of Finite Differences*, Macmillan, London.

Mitchell, A. R. [1959]: The influence of critical boundary conditions on finite difference solutions of two-point boundary value problems. *Math. Tables Aids Comput.*, **13**, 252–260.

——— and T. W. Craggs [1953]: Stability of difference relations in the solution of ordinary differential equations. *Math. Tables Aids Comput.*, **7**, 127–129.

Mohr, E. [1951]: Ueber das Verfahren von Adams zur Integration gewöhnlicher Differentialgleichungen. *Math. Nachr.*, **5**, 209–218.

Moore, R. E. [1960]: Non-linear two point boundary problems with applications to the restricted three-body problem. *Rept. LMSD*-895025, Lockheed Missiles and Space Division, Sunnyvale, Calif.

Morel, H. [1956]: Evaluation de l'erreur sur un pas dans la méthode de Runge Kutta. *C. R. Acad. Sci., Paris*, **243**, 1999–2002.

Morrison, D., and L. Stoller [1958]: A method for the numerical integration of ordinary differential equations. *Math. Tables Aids Comput.*, **12**, 269–272.

Moulton, F. R. [1926]: *New Methods in Exterior Ballistics*, Univ. Chicago Press.

Muhin, I. S. [1952a]: Application of the Markov-Hermite interpolation polynomials for the numerical integration of ordinary differential equations. *Akad. Nauk SSSR Prikl. Mat. Meh.*, **16**, 231–238.

——— [1952b]: On the accumulation of errors in numerical integration of differential equations. *Akad. Nauk SSSR Prikl. Mat. Meh.*, **16**, 753–755.

Murray, F. J. [1950]: Planning and error considerations for the numerical solution of a system of differential equations on a sequence calculator. *Math. Tables Aids Comput.*, **4**, 133–144.

*National Bureau of Standards [1942]: *Tables of Probability Functions*, New York.

*Nörlund, N. E. [1924]: *Vorlesungen über Differenzenrechnung*, Springer, Berlin.

Noumerov, B. V. [1924]: A method of extrapolation of perturbations. *Roy. Ast. Soc. Monthly Notices*, **84**, 592–601.

Nyström, E. J. [1925]: Ueber die numerische Integration von Differentialgleichungen. *Acta Soc. Sci. Fenn.*, **50**, No. 13, 1–55.

Obrechkoff, N. [1942]: Sur les quadratures méchaniques. *Spisanie Bulgar. Akad. Nauk*, **65**, 191–289.

Oldehoeft, A. E. [1959]: A numerical example in which Euler's method is stable and Milne's method is unstable. *Math. Note No.* 2, Computation Center, Univ. of North Carolina.

——— [1960]: The propagation of error in a predictor-corrector method for the solution of $dy/dx = f(x, y)$. *Mathematical Note No.* 5, Computing Center, Univ. of North Carolina.

Papoulis, A. [1952]: On the accumulation of errors in the numerical solution of differential equations. *J. Appl. Phys.*, **23**, 173–175.

Pope, D. A. [1960]: A method of "alternating corrections" for the numerical solution of two-point boundary value problems. *Math. Comput.*, 14, 354–361.

Quade, W. [1957]: Numerische Integration von gewöhnlichen Differentialgleichungen durch Interpolation nach Hermite. *Z. angew. Math. Mech.*, 37, 161–169.

——— [1959]: Ueber die Stabilität numerischer Methoden zur Integration gewöhnlicher Differentialgleichungen erster Ordnung. *Z. angew. Math. Mech.*, 39, 117–134.

Rademacher, H. [1948]: On the accumulation of errors in processes of integration on high-speed calculating machines. Proceedings of a Symposium on Large-Scale Digital Calculating machinery. *Annals Comput. Labor. Harvard Univ.*, 16, 176–187.

Ralston, A. [1960]: Numerical integration methods for the solution of ordinary differential equations, in *Mathematical Methods for Digital Computers*, A. Ralston and H. Wilf (eds.), Wiley, New York, pp. 95–109.

Rapoport, I. M. [1952]: A new method of approximate integration of ordinary differential equations. *Ukrain. Mat. Z.*, 4, 399–413.

Rice, J. R. [1960]: Split Runge-Kutta methods for simultaneous equations. *J. Research Nat. Bur. Standards*, 64B, 151–170.

Richardson, L. F. [1927]: The deferred approach to the limit, I—Single lattice. *Trans. Roy. Soc. London*, 226, 299–349.

Richter, W. [1951]: Sur l'erreur commise dans la méthode de Milne. *C. R. Acad. Sci. Paris*, 233, 1342–1344.

——— [1952]: Estimation de l'erreur commise dans la méthode de M. W. E. Milne pour l'intégration d'un système de n équations différentielles du premier ordre. Thesis, Univ. Neuchâtel.

Richtmyer, R. D. [1957]: *Difference Methods for Initial-Value Problems*, Interscience, New York.

Ridley, E. C. [1957]: A numerical method of solving second-order linear differential equations with two-point boundary conditions. *Proc. Cambridge Philos. Soc.*, 53, 442–447.

Robertson, H. H. [1960]: Some new formulae for the numerical integration of ordinary differential equations. *Information Processing*, UNESCO, Paris, pp. 106–108.

Romanelli, M. J. [1960]: Runge-Kutta methods for the solution of ordinary differential equations, in *Mathematical Methods for Digital Computers*, A. Ralston and H. Wilf (eds.), Wiley, New York, pp. 110–120.

Rothe, E. [1930]: Zweidimensionale parabolische Randwertaufgaben als Grenzfall eindimensionaler Randwertaufgaben. *Math. Ann.*, 102, 650–670.

Runge, C. [1895]: Ueber die numerische Auflösung von Differentialgleichungen. *Math. Ann.*, 46, 167–178.

Rutishauser, H. [1952]: Ueber die Instabilität von Methoden zur Integration gewöhnlicher Differentialgleichungen. *Z. angew. Math. Physik*, 3, 65–74.

——— [1955]: Bemerkungen zur numerischen Integration gewöhnlicher Differentialgleichungen n-ter Ordnung. *Z. angew. Math. Physik*, 6, 497–498.

——— [1958]: Solution of eigenvalue problems with the *LR*-transformation. *Nat. Bur. Standards Appl. Math. Ser.*, 49, 47–81.

——— [1960]: Bemerkungen zur numerischen Integration gewöhnlicher Differentialgleichungen n-ter Ordnung. *Numer. Math.*, 2, 263–279.

Salzer, H. E. [1956]: Osculatory extrapolation and a new method for numerical integration of differential equations. *J. Franklin Inst.*, 262, 111–120.

——— [1957]: Numerical integration of $y'' = \varphi(x, y, y')$ using osculatory interpolation. *J. Franklin Inst.*, 263, 401–409.

Schröder, J. [1956]: Ueber das Differenzenverfahren bei nichtlinearen Randwertaufgaben. I. *Z. angew. Math. Mech.*, 36, 319–331; II. 443–455.

——— [1958]: Fehlerabschätzungen bei gewöhnlichen und partiellen Differentialgleichungen. *Arch. Rational Mech. Anal.*, **2**, 367–392.

——— [1961a]: Fehlerabschätzungen mit Rechenanlagen bei gewöhnlichen Differentialgleichungen erster Ordnung. *Num. Math.*, **3**, 39–61.

——— [1961b]: Verbesserung einer Fehlerabschätzung für gewöhnliche Differentialgleichungen erster Ordnung. *Numer. Math.*, 3, 125–130.

Sconzo, P. [1954]: Formule d'estrapolazione per l'integrazione numerica delle equazioni differenziali ordinarie. *Boll. Un. Mat. Ital.* (3), **9**, 391–399.

Serrais, F. [1956]: Sur l'estimation des erreurs dans l'intégration numérique des équations différentielles linéaires du second ordre. *Ann. Soc. Sci. Bruxelles, Sér. I*, **70**, 5–8.

Sheldon, J. W., B. Zondek, and M. Friedman [1957]: On the time-step to be used for the computation of orbits by numerical integration. *Math. Tables Aids Comput.*, **11**, 181–189.

Shura-Bura, M. R. [1952]: Estimates of errors of numerical integration of ordinary differential equations. *Akad. Nauk SSSR Prikl. Mat. Meh.*, **16**, 575–588.

[illegible]ith, O. K. [1959]: The inverse matrix of solutions of the variational equations. *Tech. Note PA-1951-16/2*, Space Technology Laboratories, Los Angeles.

Stafford, H. R. [1959]: On linear differential and difference equations. M.A. Thesis, Univ. of Kansas.

Sterne, T. E. [1953]: The accuracy of numerical solutions of ordinary differential equations. *Math. Tables Aids Comput.*, **7**, 159–164.

Stohler, [illegible] [1943]: Eine Vereinfachung bei der numerischen Integration gewöhnlicher Differentialgleichungen. *Z. angew. Math. Mech.*, **23**, 120–122.

Störmer, C. [1907]: Sur les trajectoires des corpuscules électrisés. *Arch. sci. phys. nat., Genève*, **24**, 5–18, 113–158, 221–247.

——— [1921]: Méthodes d'intégration numérique des équations différentielles ordinaires. *C. R. congr. intern. math., Strasbourg*, pp. 243–257.

Stüssi, F. [1950]: Numerische Lösung von Randwertproblemen mit Hilfe der Seilpolygongleichung. *Z. angew. Math. Mech.*, **1**, 53–70.

*Taylor, A. E. [1955]: *Advanced Calculus*, Ginn, New York.

Tihonov, A. N., and A. A. Samarskii [1956]: On finite difference schemes for equations with discontinuous coefficients. *Dokl. Akad. Nauk SSSR* (N. S.), **108**, 393–396.

Todd, J. [1950]: Notes on numerical analysis, I. Solution of differential equations by recurrence relations. *Math. Tables Aids Comput.*, **4**, 39–44.

Tollmien, W. [1938]: Ueber die Fehlerabschätzung beim Adamsschen Verfahren zur Integration gewöhnlicher Differentialgleichungen. *Z. angew. Math. Mech.*, **18**, 83–90.

——— [1953]: Bemerkung zur Fehlerabschätzung beim Adamsschen Interpolationsverfahren. *Z. angew. Math. Mech.*, **33**, 151–155.

Uhlmann, W. [1957a]: Fehlerabschätzungen beim Anfangswertproblem gewöhnlicher Differentialgleichungssysteme 1. Ordnung. *Z. angew. Math. Mech.*, 37, 88–99.

——— [1957b]: Fehlerabschätzung bei Anfangswertaufgaben einer gewöhnlichen Differentialgleichung höherer Ordnung. *Z. angew. Math. Mech.*, **37**, 99–111.

Urabe, M. [1960]: Theory of errors in numerical integration of ordinary differential equations. *Tech. Summary Rept. No. 183*, U.S. Army Mathematics Research Center, Madison, Wis.

——— [1961]: Theory of errors in numerical integration of ordinary differential equations. *J. Sci. Hiroshima Univ., Ser.* A-1, **25**, 3–62.

——— and S. Mise [1955]: A method of numerical integration of analytic differential equations. *J. Sci. Hiroshima Univ. Ser. A*, **19**, 307–320.

Urabe, M., and T. Tsushima [1953]: On numerical integration of ordinary differential equations. *J. Sci. Hiroshima Univ. Ser. A*, **17**, 193–219.

Urabe, M., and H. Yanagihara [1954]: On numerical integration of the differential equation $y^{(n)} = f(x, y)$. *J. Sci. Hiroshima Univ. Ser. A*, **18**, 55–76.

*van der Waerden, B. [1950]: *Modern Algebra*, Vol. 1, translated by T. J. Benac, F. Ungar Pub. Co., New York.

van Wijngaarden, A. [1953]: Erreurs d'arrondiment dans les calculs systématiques. Les machines à calculer et la pensée humaine, *Colloq. intern. centre nat. recherche sci., Paris*, no. 37, pp. 285–293.

Varga, R. S. [1959]: *Iterative Numerical Analysis*, Carnegie Institute of Technology, Pittsburgh.

Vietoris, L. [1953a]: Der Richtungsfehler einer durch das Adamssche Interpolations-verfahren gewonnenen Näherungslösung einer Gleichung $y' = f(x, y)$. *Oesterr. Akad. Wiss., Math.-naturw. Kl., Sitzber., Abt. IIa*, **162**, 157–167.

——— [1953b]: Der Richtungsfehler einer durch das Adamssche Interpolations-verfahren gewonnenen Näherungslösung eines Systems von Gleichungen $y_k' = f_k(x, y_1, \cdots, y_m)$. *Oesterr. Akad. Wiss. Math.-naturw. Kl., Abt. IIa*, **162**, 293–299.

Vlasov, I. O., and I. A. Čarnyi [1950]: On a method of numerical integration of ordinary differential equations. *Akad. Nauk SSSR Inženernyĭ Sbornik*, **8**, 181–186.

von Mises, R. [1930]: Zur numerischen Integration von Differentialgleichungen. *Z. angew. Math. Mech.*, **10**, 81–92.

von Oppolzer, T. R. [1880]: *Lehrbuch zur Bahnbestimmung der Kometen und Planeten*, Vol. 2, W. Engelmann, Leipzig.

von Sanden, H. [1945]: *Praxis der Differentialgleichungen*, 3rd. ed., de Gruyter, Berlin.

Vorobev, L. M. [1956]: The applicability of S. A. Caplygin's method of approximate integration to a certain class of ordinary nonlinear differential equations of the second order. *Uspehi Mat. Nauk* (N. S.), **11**, 181–185.

Voronovskaya, E. V. [1955]: On an alteration of Caplygin's method for differential equations of the first order. *Prikl. Mat. Meh.*, **19**, 121–126.

Wachspress, E. L. [1960]: The numerical solution of boundary value problems, in *Mathematical Methods for Digital Computers*, A. Ralston and H. Wilf (eds.), Wiley, New York. pp. 121–127.

Wall, D. D. [1956]: Note on predictor-corrector formulas. *Math. Tables Aids Comput.*, **10**, 167.

Warga, J. [1953]: On a class of iterative procedures for solving normal systems of ordinary differential equations. *J. Math. Physics*, **31**, 223–243.

Wasow, W. [1955]: Discrete approximations to elliptic differential equations. *Z. angew. Math. Physik*, **6**, 81–97.

Weissinger, J. [1950]: Eine verschärfte Fehlerabschätzung zum Extrapolationsverfahren von Adams. *Z. angew. Math. Mech.*, **30**, 356–363.

——— [1952]: Eine Fehlerabschätzung für die Verfahren von Adams und Störmer. *Z. angew. Math. Mech.*, **32**, 62–67.

——— [1953]: Numerische Integration impliziter Differentialgleichungen. *Z. angew. Math. Mech.*, **33**, 63–65.

Wilf, H. S. [1957]: An open formula for the numerical integration of first-order differential equations. *Math. Tables Aids Comput.*, **11**, 201–203.

——— [1959]: A stability criterion for numerical integration. *J. Assoc. Comput. Mach.*, **6**, 363–365.

——— [1960]: Maximally stable numerical integration. *J. Soc. Ind. Appl. Math.*, **8**, 537–540.

Young, D. M. [1955]: Gill's method for solving ordinary differential equations. *Numerical Note N-N 4*, Ramo-Wooldridge Corp., Los Angeles.
——— [1957]: Review of Wall [1956]. *Math. Rev.*, 18, 336.
Zadiraka, K. V. [1951]: Solution by the method of S. A. Caplygin of linear differential equations of the 2nd order with variable coefficients. *Dopovidi Akad. Nauk Ukrain. RSR*, 1951, 163–170.
——— [1952]: The approximate integration by S. A. Caplygin's method of linear differential equations of the 2nd order with variable coefficients. *Ukrain. Mat. Z.*, 4, 299–311.
——— [1955]: The construction of upper and lower bounds for the eigenvalues of one-dimensional self-adjoint boundary value problems of even order. *Doklady Akad. Nauk SSSR*, 102, 681–684.
——— and I. B. Pogrebiskii [1950]: On the application of S. A. Caplygin's method of approximate integration of ordinary differential equations to a boundary problem. *Dopovidi Akad. Nauk Ukrain. RSR*, 1950, 95–100.
Zondek, B., and J. W. Sheldon [1959]: On the error propagation in Adams' extrapolation method. *Math. Tables Aids Comput.*, 13, 52–55.
Zurmühl, R. [1948]: Runge-Kutta Verfahren zur numerischen Integration von Differentialgleichungen *n*-ter Ordnung. *Z. Angew. Math. Mech.*, 28, 173–182.
——— [1952]: Runge-Kutta Verfahren unter Verwendung höherer Ableitungen. *Z. angew. Math. Mech.*, 32, 153–154.